Cytometry
PART A
Journal of the International Society for Advancement of Cytometry

Large-Scale Image-Based Screening and Profiling of Cellular Phenotypes

Nicola Bougen-Zhukov,[1] Sheng Yang Loh,[1] Hwee Kuan Lee,[1] Lit-Hsin Loo[1,2]*

[1]Bioinformatics Institute, Agency for Science, Technology and Research, Singapore 138671, Singapore

[2]Department of Pharmacology, School of Medicine, National University of Singapore, Singapore 117600, Singapore

Grant sponsor: Joint Council Office, Grant number: 1431AFG105

Grant sponsor: Bioinformatics Institute

Correspondence to: Bioinformatics Institute, 30 Biopolis Street, #07-01 Matrix, Singapore 138671, Singapore. E-mail loolh@bii.a-star.edu.sg

Published online 19 July 2016 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.22909

• **Abstract**

Cellular phenotypes are observable characteristics of cells resulting from the interactions of intrinsic and extrinsic chemical or biochemical factors. Image-based phenotypic screens under large numbers of basal or perturbed conditions can be used to study the influences of these factors on cellular phenotypes. Hundreds to thousands of phenotypic descriptors can also be quantified from the images of cells under each of these experimental conditions. Therefore, huge amounts of data can be generated, and the analysis of these data has become a major bottleneck in large-scale phenotypic screens. Here, we review current experimental and computational methods for large-scale image-based phenotypic screens. Our focus is on phenotypic profiling, a computational procedure for constructing quantitative and compact representations of cellular phenotypes based on the images collected in these screens.

• **Key terms**

high-content screening; phenotypic profiling; imaging-based phenotypic screens; automated image analysis; cellular phenotypes; high-throughput microscopy

Introduction

Cellular phenotypes are observable characteristics of cells resulting from the interactions of intrinsic and extrinsic factors (Fig. 1). Intrinsic factors include biomolecules, such as DNA, RNA, proteins, or metabolites, produced within the cells. Extrinsic factors include biomolecules or chemicals that originate from extracellular sources, such as other cells, the environment, or man-made sources. Due to the advances in automated microscopy and image analysis, it has become feasible to image cellular phenotypes under large numbers of experimental conditions that mimic the influences of these intrinsic or extrinsic factors. Often, such screens are performed by varying either the intrinsic or extrinsic factors, while keeping all other factors or conditions unchanged (Fig. 1). "Intrinsic-phenotype" screens can be used to study phenotypes by monitoring different intrinsic factors while keeping cells under the same extrinsic factors or environmental conditions. Alternatively, "extrinsic-phenotype" screens can be used to study phenotypes by subjecting cells to different extrinsic factors while monitoring the same intrinsic biomolecular species. Although the purposes of intrinsic- and extrinsic-phenotype screens are different, they often use analogous experimental and computational methods (Fig. 1). These screens can generate images for millions of cells, from each of which thousands of numerical descriptors (or "features") of phenotypes can be measured. These huge amounts of generated data, which are usually in high dimensions, are challenging for manual analysis. This problem has become a major bottleneck in large-scale image-based phenotypic screens. Here, we review current experimental and computational methods for large-scale image-based phenotypic screens. We show how intrinsic- and extrinsic-phenotype screens can produce vast amounts of images, and discuss a

ISAC
International Society for Advancement of Cytometry

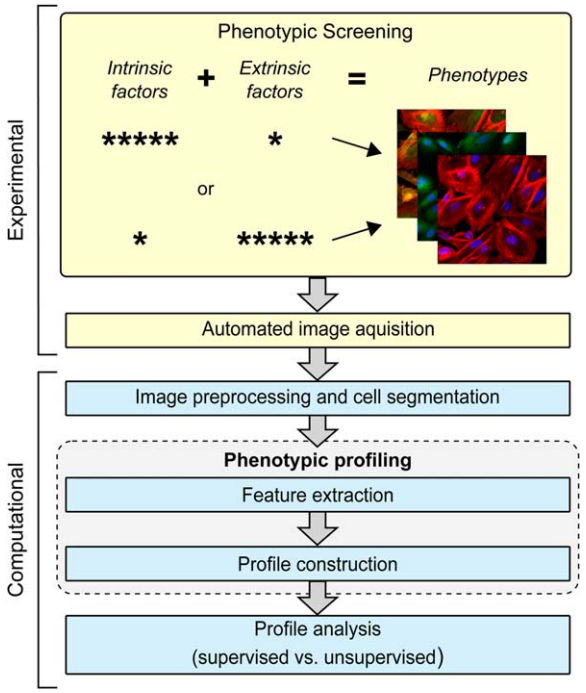

Figure 1. Experimental and computational workflow of large-scale image-based phenotypic screening. Intrinsic-phenotype screens involve the screening of many intrinsic factors (such as RNAs, proteins, or metabolites), while keeping the extrinsic factors or conditions constant; whereas extrinsic-phenotype screens involve the screening of many extrinsic factors (such as synthetic small molecules or oligonucleotides), while monitoring the same intrinsic factors. [Color figure can be viewed at wileyonlinelibrary.com]

general computational workflow for analyzing these images. Our focus is on phenotypic profiling, a computational procedure to build more compact representations (or "profiles") of the data generated from these screens, while keeping most of the biological information intact. We also discuss the current challenges in the adoption and use of phenotypic profiling in large-scale phenotypic screens.

Image-Based Phenotypic Screens

Intrinsic Phenotypes

Intrinsic-phenotype screens are often used to infer the biological functions of novel or uncharacterized intrinsic biomolecules (1–3). This is because biomolecules that perform similar biological functions or participate in related biological processes tend to express similar phenotypes, such as subcellular localization patterns (2,3), interaction partners (4), or expression variations (5). To monitor the phenotypes of large numbers of biomolecular species, scalable methods for labeling biomolecules are required. Among all the biomolecules, the techniques for labeling proteins, one of the fundamental building blocks of cells, are the most developed and widely used. Endogenous proteins can be labeled using primary antibodies followed by amplification with secondary antibodies that are conjugated to fluorescent dyes (6). The Human Protein Atlas project has generated a repository of >25,000 specific antibodies, targeting proteins from >17,000 human genes (7). Immunohistochemistry and immunofluorescence images of tissues or cells stained with most of these antibodies are available on the web portal of the project (http://www.proteinatlas.org). Alternatively, proteins can be covalently labeled with genetically encoded fluorescent-protein (FP) fusion tags (6). Large-scale FP tagging has been performed for hundreds to thousands of genes in the budding yeast *Sacchramyces cerevisiae* (2), the fruit fly *Drosophila* (8), and human cells (9). This labeling method allows the study of dynamic cellular phenotypes using live-cell imaging; therefore, it can generate tremendously more numbers of images than antibody-based methods.

Labeling RNAs is more challenging than labeling proteins, partly due to the shorter half-lives and lower abundances of RNA molecules in the cells. For example in the budding yeast, it was estimated that ~4,000–6,000 protein molecules are translated per mRNA molecule (10,11). Fluorescence in situ hybridization (FISH) can be used to fluorescently label and image RNAs via the hybridization of sequence-specific fluorescent oligonucleotides (12). The signal intensity of FISH can be increased by using multiple short oligonucleotide probes to adjacent sequences on a RNA target (13), and/or conjugating multiple fluorescent dyes to each oligonucleotide probe (14). Then, single mRNA molecules can be detected as diffraction-limited spots under fluorescence microscopy. This methodology is called single-molecule FISH (smFISH). It can now be used to image the RNAs for ~1,000 genes in different (15) or even the same (16) single human cells, thereby enabling image-based single-cell transcriptomic screens. The traditional FISH method has also been optimized to allow the imaging of tens to thousands of mRNAs in neurons (17) and the *Drosophila* embryos (18). With the advent of all these biomolecule labeling techniques, we can now image and measure the abundance and localization of large numbers of intrinsic biomolecular species at the single-cell level.

Despite the enduring doctrine that RNA expression leads to protein expression, the relationship between RNA and protein expressions are not always positively correlated. A previous study of the NCI-60 human cancer cell lines found that only ~65% of the genes showed significant gene-protein expression correlations (19). A more recent study, which measured mRNA and protein expression levels in parallel from the same cell populations, shows that mRNA levels explain only ~40% of the variability in protein levels (20). Therefore, intrinsic phenotypic screens based on RNA or protein labeling may convey different information. The former may be more useful for studying transcription-driven processes, such as cell differentiation (18) or division; while the latter may be more useful for studying protein-modification or translocation-driven processes, such as drug response (21,22) or vesicle transport (23).

However, the use of RNA or protein labeling in phenotypic screens is often limited by other more practical considerations. Proteome-wide screening using antibodies is expensive and time consuming. Furthermore, antibodies for specific isoforms or modified proteins may be hard to obtain or generate.

Although the design and synthesis of RNA oligonucleotide probes are relatively easy and in theory can cover all possible transcripts, certain RNA species may be hard to detect simply due to their lower abundances in the cells. Finally, live-cell imaging of RNA and protein molecules remains very challenging. Transient overexpression of RNAs fused with RNA-aptamer-fluorophore complexes (24) or proteins fused with FPs (6) may be used, but these overexpression techniques may interfere with the functions of endogenous RNAs or proteins and/or cause undesirable cellular toxicity (25). Several new labeling techniques have been recently developed to address some of these problems. They include genome-editing techniques to express FP tags at near endogenous levels by integrating them directly into native genomic loci (26) or generate programmable nuclease complexes that can recognize different endogenous mRNAs (27). Low-toxicity chemical probes that can label endogenous DNA, actin cytoskeleton, or microtubules in live cells have also been developed (28). These exciting techniques may make large-scale intrinsic-phenotype screens in live cells more efficient and feasible.

Extrinsic Phenotypes

Extrinsic-phenotype screens are often used to identify extrinsic factors or perturbations that can induce certain desired cellular phenotypes. These screens differ from intrinsic-phenotype screens, in which different extrinsic factors or perturbations are applied to the cells, while the same biomolecular species are monitored (Fig. 1). Synthetic or natural chemical libraries are examples of extrinsic perturbations that have been widely used for large-scale screening. The development of large libraries of structurally diverse chemicals is mainly driven by the needs of the pharmaceutical industry in drug discovery (29). Large drug screening programs in pharmaceutical companies can generate $>$50 million data points per screening campaign, where each of the data points is from a compound tested at a concentration (30). A survey of the new drugs approved between 1999 and 2008 found that most of the first-in-class small-molecule drugs were discovered through phenotypic screening (31). However, cellular imaging is a relatively new approach for phenotypic screens. Only since the turn of the millennium has image-based approach steadily gained traction (32), and been used to screen for small-molecule inhibitors of signal transduction (33,34), viral infection (35), and blood-vessel angiogenesis (36).

The mechanisms of action of extrinsic perturbations are often unknown. Such are the cases for novel or uncharacterized synthetic small molecules, natural products, or environmental toxicants. However, extrinsic factors with similar targets or mechanisms are likely to induce similar changes in cellular phenotypes (22,37). Therefore, extrinsic-phenotype screens can still be used to associate these factors together and separate them from other extrinsic factors that have different mechanisms or modes of action. In many applications, such as drug discovery, the molecular targets of the selected extrinsic factors (or "hits") will need to be further identified, so that the hits can be chemically optimized. This type of phenotypic screens is also called "forward chemical-genetic" screens (29). For other applications, such as toxicity classification (38–40), the tested extrinsic factors may have multiple, diverse, or even non-specific intracellular targets. Sometimes, these extrinsic factors, such as environmental pollutants or natural-product extract samples, may even consist of mixtures of chemicals with unknown proportions and identities (39). Phenotypic screening is one of the few viable approaches for characterizing or classifying these types of extrinsic perturbations (40).

If the molecular targets of the extrinsic perturbations are known and specific, extrinsic-phenotype screens can be used to directly infer the biomolecules that are involved in the generation of the observed phenotypes. This type of phenotypic screen is also called "reverse genetic or chemical-genetic" screens (29). RNA interference (RNAi) and the clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated-protein-9 (Cas9) system are two genetic perturbation techniques based on the cellular adaptive immune systems against viruses or other foreign genetic materials. RNAi is a form of post-transcriptional gene regulation, in which long double-stranded RNA (dsRNA) molecules of exogenous or endogenous origins are cleaved into small interfering RNA (siRNA) molecules that mediate sequence-specific degradation of messenger RNA (mRNA) molecules (41). The CRISPR/Cas9 system is a prokaryotic immune system that can recognize and degrade foreign double-stranded DNA from bacteria, viruses, or plasmids through the introduction of double stranded breaks (42). Although RNAi and CRISPR/Cas9 were first discovered in plants and bacteria, respectively, they can also be used to knockdown, knockout, or even enhance the expressions of specific genes in other species, including human, at the cellular (43–45) or organismal (46,47) levels. Genome-wide siRNA or CRISPR-Cas9 libraries for human genes have been developed (48,49). Image-based phenotypic screens based on these or similar libraries have been used to identify genes involved in cell division (50), cell migration (51), endocytosis (52), and chromosome segregation (53).

One of the main advantages of using a CRISPR/Cas9 system is that complete knockouts of genes can be achieved, thereby enabling direct inferences of the relationships between genotypes and phenotypes. However, the knockouts of certain genes, especially essential genes, may result in cell death or other severe defects that can mask the phenotypes of interest. RNAi, which only reduces the expression of genes, may still be used to study the phenotypes associated with these genes. Interestingly, a recent comparative study found that large-scale CRISPR/Cas9- or RNAi-based phenotypic screens identify gene sets that have little correlation and different enrichments of biological processes (54). Therefore, these genetic perturbation techniques are not directly interchangeable, and more studies are still required to better understand their genomic coverage and selectivity. The usage of these techniques is also limited by several practical considerations. Most of them require a multi-day transfection or delivery stage, which is both time-consuming and artifact-prone. For example, variability in transfection may lead to variability in knockdown efficiency (55), and the transfection process itself may trigger

unwanted cellular responses (56). For certain phenotypic screens, especially those aim to identify extrinsic factors that can rapidly inhibit protein post-translational modifications, perturbations based on small molecules may still be more desirable due to their rapid (often within minutes) and efficient uptake by cells. There is also an increasing interest to screen or design polypharmacological compounds that can inhibit multiple intra-cellular targets, which may lead to more effective but less toxic therapeutic agents (57). Therefore, despite the difficulty in de-convoluting the targets of small molecules, they remain one of the most commonly used perturbation methods in extrinsic-phenotype screens.

Automated Image Acquisition and Processing

Image Acquisition

Imaging is one of the most direct ways to observe cellular phenotypes. Technological advances in automated microscopy have enabled near-autonomous large-scale image-based phenotypic screens. The developments of robotics for sample preparation, liquid dispensing, plate handling, and microscope control have enabled the imaging of cells under huge numbers of diverse intrinsic or extrinsic factors (53,58,59). Although most of the experimental processes can be automated, imaging large numbers of cells still requires considerable time. Focus control is one of the bottlenecks in microscope automation (60). Image-based autofocus methods are slow, and have been mostly replaced with faster and more robust reflection-based autofocusing methods (61), which are commercially available from several vendors of microscopes. However, most of the reflection-based methods work by maintaining constant vertical offsets between the imaging interfaces and objective lens, and may still have problems imaging samples with uneven thickness (either due to the culture/supporting substrates or the cell/tissue specimens themselves). To reduce imaging time, especially for live-cell imaging, machine-learning-based methods and tools, such as the Micropilot platform (62), can be used to recognize, track, and image only certain selected subset of cells.

Assay miniaturization has also contributed to the increase in the scale of phenotypic screens. There are currently two approaches for assay miniaturization. The first approach involves microtiter plates with higher well density and smaller well volume, such as the 384- or even 1536-well plates (37,53,63). The second approach uses microfluidic devices, which are miniaturized fluidic channels (usually <1 mm) that can perform liquid handling and perturbation experiments (64,65). All of these technologies reduce the amount of biological and chemical reagents used per assay, and therefore allow more phenotypic screening experiments to be performed.

Image Preprocessing and Cell Segmentation

Due to the large numbers of generated images, automated image processing is required and critical for large-scale image-based phenotypic screen. First, image pre-processing algorithms are used to reduce noise and correct non-even background or illumination in the images (66). For certain applications, image stitching is also used to combine images taken from multiple positions (67). Then, automated cell segmentation is used to identify cellular or subcellular regions from the acquired images (68). The accuracy of cell segmentation affects the subsequent measurements of cellular phenotypes, especially for morphological properties—such as cell size and roundness. Cell segmentation algorithms usually detect cellular regions based the staining properties, such as brightness, size, and gradient, of fluorescence labels that specifically stain the entire cells or only the nucleus. Common segmentation algorithms include watershed transformation and level-set methods (68). There are also algorithms that are optimized for specific applications, such as segmentation of touching cells (69–71), overlapping cells (72), texture-based regions (73), and protein aggregates (74). Readers may refer to other previous reviews for more information about image pre-processing (66) and cell segmentation (68).

Feature Extraction

After identifying the image regions that correspond to individual cells, numerical phenotypic descriptors (or "features") can be used to quantify changes in the intrinsic or extrinsic phenotypes depicted in these regions. This process is called "feature extraction." The advent of more powerful computers has enabled the automation of feature extraction. As early as in the 1960s, automated measurements of the length, size, skewness, and other morphological properties of chromosomes (75) and neurons (76) from microscopy images had been performed. Most of the current commonly used phenotypic features are "location-independent" features, whose values are based on either the statistics of the pixel values in the cellular regions, or the shape properties of the outlines of the regions (Fig. 2). These features are location-independent because they do not consider the locations of individual pixels within the cellular regions, and would give exactly the same values even if the positions of the underlying pixels are randomly shuffled. Examples of this type of features include cellular morphology (e.g., cell size and roundness), organelle structures (e.g., mitochondria or nuclear sizes), and intracellular levels of biomolecules (e.g., mean intensity values of the fluorescent labels for these molecules) (Fig. 2). The usage of "location-dependent" features is rare, except for the measurements of the localizations of biomolecules at specific subcellular regions (e.g., nuclear-to-cytoplasmic intensity ratios of the fluorescent labels for these molecules). Most of the current phenotypic features are manually designed or selected to be readouts for specific biological processes or phenotypes of interest, and thus they usually have direct biological interpretations. For example, blebbing of the nucleus (77) and DNA fragmentation (78) are indicators of cell death, and nuclear translocation of transcription factors often reflects activation of transcription (79). Due to the rich information that can be extracted from cellular images, image-based phenotypic screens are also commonly referred to as "high-content" screens (80). Several freely available biological image-processing software packages, including ImageJ/Fiji (81), OMERO (82), CellProfiler (83), and cellXpress (84), can be used to perform automated cell segmentation and feature extraction (Table 1).

Image-Based Phenotypic Profiling

New Challenges in Large-Scale Screens

The advent of large-scale screening has created new challenges to the feature extraction and analysis steps in image-based phenotypic screens. Due to the large numbers of experimental conditions tested in these screens, changes in the intrinsic or extrinsic phenotypes are usually unknown before the experiments, or manifested in complex or non-continuous forms. For example, a previous genome-wide siRNA screen induced at least 16 distinct morphological changes in the nuclei of cells during cell division (50). This creates two new problems. First, it is difficult to specifically design quantitative features for these phenotypes *a priori*. There is an increasing awareness in the community that most current high-content studies actually rely on only one or two manually selected phenotypic features (32), most of which are location-independent features as discussed above. The usage of these features is prone to human selection bias and may not lead to the discovery of novel phenotypes. Second, the numbers and definitions of all possible categories of phenotypic changes are usually unknown. This prohibits the use of supervised machine-learning methods, which automatically learn and build computational models based on labeled training data that represent pre-defined categories of phenotypes (85). These two problems were avoided in most of the past studies by using manually defined and/or assigned categories of phenotypes. For example, several previous intrinsic-phenotype screens, such as the original genome-wide yeast protein localization study (2) and Human Protein Atlas project (7), used pre-defined categories that correspond to known subcellular compartments or organelles. The assignment of proteins to these categories were also performed manually (2). Similarly, several extrinsic-phenotype screens based on genome-wide siRNA libraries also classified the resulting phenotypes into manually defined phenotypic categories (50,52,86). Although these categories are human interpretable and allow the usage of supervised data analysis methods, they provide limited representations of the rich information contained in the images. For example, manual assignments based on visual inspections may have difficulty in distinguishing proteins that localize in multiple subcellular compartments or in different ratios (3). Extrinsic factors may also induce new phenotypes beyond these pre-defined categories and these changes may be completely missed. Even for well-defined phenotypes, such as the chromatin morphology for different stages of mitosis, user annotations were found to be inconsistent (87). Therefore,

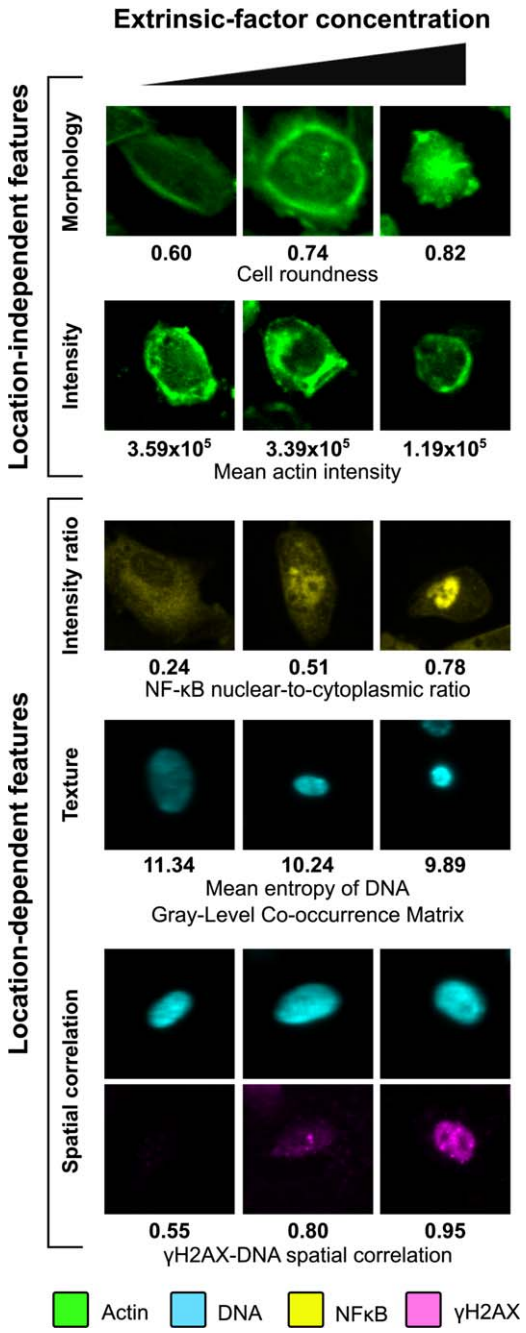

Figure 2. Types of phenotypic features that can be quantified from microscopy images of cells. Location-independent features can be based on the shape properties of the outlines of cellular regions (e.g., cell roundness), or the statistics of the intensity values of the pixels within the regions (e.g., mean intensity value of the staining of a cytoskeleton marker, actin, in the cellular region). These features are location-independent because they would give the same values even if the positions of the pixels within the same regions are randomly shuffled. Location-dependent features can be based on the intensity ratios of the same biomolecule markers at different subcellular regions [e.g., the nuclear-to-cytoplasmic ratio of a transcription factor, nuclear factor kappa B (NF-κB)], textural properties of the marker staining patterns [e.g., mean entropy of the gray-level co-occurrence matrix (90) based on a DNA marker], and spatial correlations between two different markers (e.g., correlation coefficient between the markers for DNA and phosphorylated histone 2A-X, γH2AX, at the cellular regions). Examples of primary human proximal tubular cells with different values of these features are shown. The cells were treated with nephrotoxic compounds (extrinsic factors) at different concentrations, which induce the changes in cellular phenotypes. More information about the cells, features, and experimental protocols for generating the shown images can be found in our previous report (40). [Color figure can be viewed at wileyonlinelibrary.com]

Table 1. Freely available phenotypic screening and profiling software tools

	SOFTWARE DESIGN			PHENOTYPIC PROFILING					
	GUI	MULTI-CPU SUPPORT	MULTI-WELL PLATE BROWSER	INTENSITY FEATURES	SHAPE FEATURES	TEXTURE FEATURES	OBJECT-BASED FEATURES	PROFILE CONSTRUCTION	PROJECT WEBSITES
CecogAnalyzer (1.6)	+	−	−	+	+	+	−	+	www.cellcognition.org
CellProfiler (2.1.1)/ CellProfiler Analyst (2.0)	+	+	+	+	+	+	+	+	www.cellprofiler.org
cellXpress (1.3)	+	+	+	+	+	+	+	+	www.cellxpress.org
EBImage (4.12.2)	−	−	−	+	+	+	+	−	www.bioconductor.org
Icy (1.7.3)	+	+	−	+	+	−	+	−	icy.bioimageanalysis.org
Image J (1.50)	+	+	−	+	+	−	+	−	imagej.nih.gov/ij
OMERO (5.2)/ WND-CHARM	+	+	+	+	+	+	+	+	www.openmicroscopy.org

new automated data analysis methods and tools are needed to characterize the phenotypic changes in large-scale screens.

Phenotypic Profiling

Phenotypic profiling is a procedure to construct quantitative representations (or "profiles") of cellular phenotypes based on the images collected in large-scale phenotypic screens (Fig. 3). These profiles are usually used to build computational models that can automatically classify or group intrinsic or extrinsic factors in the screens. The idea of phenotypic profiling was first demonstrated by Murphy and colleagues. A set of 84 image features were measured from cellular images, and then stepwise discriminant analysis was used to identify a subset of 37 features that could classify proteins localized in ten different subcellular compartments (88). Since then, several other studies have demonstrated that phenotypic profiling may also be used to classify the effects of small molecules (22,34,37), identify novel biomolecules that mediate biological processes (50,52,86), annotate protein localization patterns (3,89), compare spatial and functional divergence of proteins (3), or predict the toxicity of xenobiotic compounds (40).

There are two unique characteristics that distinguish phenotype profiling from other high-content analysis (HCA) methods (Fig. 3). The first characteristic is that large numbers of general phenotypic features are usually automatically and

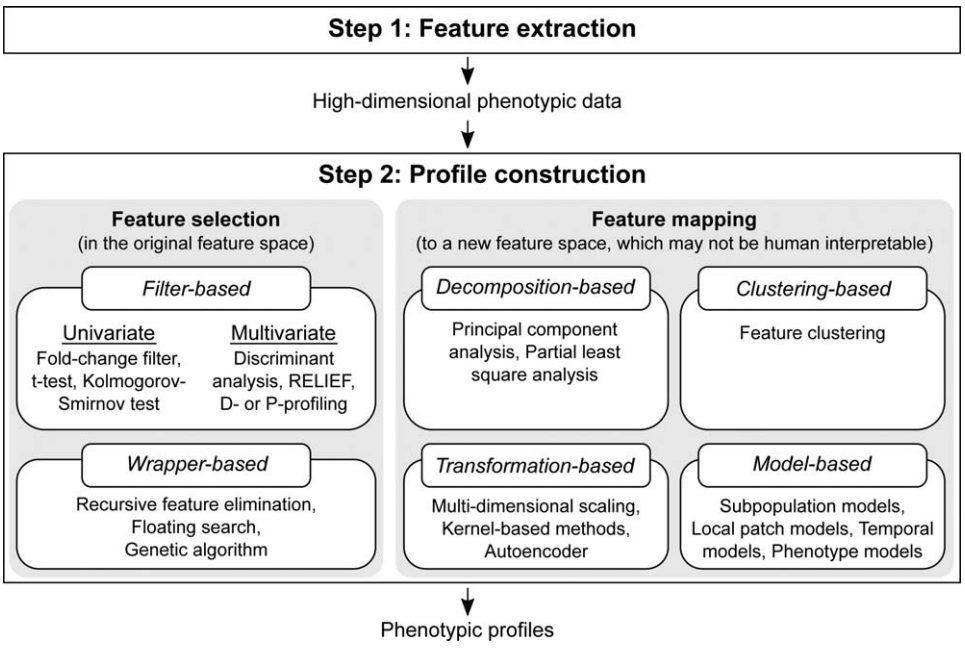

Figure 3. Computational workflow of phenotypic profiling. The first step of phenotypic profiling is to generate large numbers of phenotypic features, which usually include many spatial-dependent features. The second step is to construct compact representations of cellular phenotypes based on the extracted high-dimensional data. Several examples of profile construction methods are shown.

unbiasedly generated from the cellular images. In addition to the location-independent features mentioned above, phenotypic profiling usually also measures many location-dependent features, such as the Haralick's (22,40,88,90,91) or Gabor (92) texture features, which describe the spatial arrangements of the intensity values of neighboring pixels in the images; moment and wavelet features (88,93), which measure the distributions of the intensity values of all pixels with respect to some axes or basis; and local structures, which measure the statistics of local objects at different scales in the images (3). These features can be used to describe complex spatial distribution patterns of intrinsic biomolecules in the cells (3,22,40,88,89). In most applications, these features are measured only in the detected cellular or subcellular regions (3,22,40), and thus they represent local properties of the images. However, the same features may also be measured globally in the entire images (88,89,94). There is also a special type of feature extraction methods called "local key-point feature detection," which identifies key points in the entire images and extracts local features in the regions around these points. Examples of such methods are the Scale-Invariant Feature Transform (SIFT) (95) and Speeded Up Robust Features (SURF) (89). Both global and local key-point features do not require cell segmentation, and thus may be potentially faster than segmentation-based features. Unlike standard HCA methods, all of the above mentioned features are not designed for any specific cellular phenotype. In theory, some of these features, such as the moment and wavelet features, can reconstruct any spatial distribution patterns of the intrinsic biomolecules (93). However, due to the non-specificity of these features, many of them are expected to be irrelevant or redundant for the representations of the observed cellular phenotypes.

The second unique characteristic of phenotypic profiling is the conversion of the measured high-dimensional feature data into compact and representative phenotypic profiles. This can be done by either selecting optimum feature subsets from all the extracted features ("feature selection"), or mapping the high-dimensional feature data to a lower-dimensional feature space ("feature mapping") (Fig. 3). Feature mapping can usually remove more redundant or irrelevant features than feature selection. However, the lower-dimensional feature space may not be human interpretable, and thus is not preferable when the understanding or interpretation of the observed phenotypes is required. Most of the feature selection and mapping methods developed for high-dimensional data in the data mining, computer vision, document classification, and bioinformatics fields are directly applicable to phenotypic screening data, but a few methods have been specifically designed for phenotypic profiling. In the following paragraph, we will provide some examples for each of these methods, and highlight those that are more recently developed or relevant to image-based phenotypic profiling. Phenotypic profiling is conceptually very similar to the approach of gene expression profiling, where the expressions of large numbers of genes are first measured and then the differentially expressed genes are automatically identified. Most conventional HCA only use small numbers of manually designed features, and therefore they do not need to construct phenotypic profiles.

Feature Selection and Mapping

Feature selection methods can be divided into two main types (96) (Fig. 3). First, filter-based methods select features regardless of the computational models that will be built. They include univariate methods, such as the fold-change filter, variants of *t* test, and Kolmogorov-Smirnov test (37); and multivariate methods, such as discriminant function analysis (88,97) and the Relief algorithm (98). Most of these methods are based on the comparisons of the statistics or properties of the feature distributions under treated and control conditions, and therefore are more widely used in extrinsic-phenotype screens. We have previously developed the Drug Profiling ("D-profiling") (22) and Protein-localization Profiling ("P-profiling") algorithms (3), which can be used to both extrinsic- and intrinsic-phenotype screens. The algorithms construct phenotypic profiles based on hyperplanes that optimally separate cells under the tested extrinsic or intrinsic factors and reference conditions in the multidimensional feature space. Most of the filter-based methods, including D- and P-profiling, do not require pre-defined categories of phenotypes, and therefore can be used even if the changes in extrinsic or intrinsic phenotypes are unknown before the experiments (3,22). Second, wrapper-based methods select features by using the eventual computational models to evaluate feature subsets. They include recursive feature elimination (22,40,99), floating search (100), and genetic algorithm (101). Wrapper-based methods usually produce more compact phenotypic profiles and more predictive computational models than filter-based methods. However, they require pre-definition of phenotypes, and cannot be used if the changes in phenotypes are unknown before the experiments. Also, they are usually more computationally intensive and take longer time to complete.

Feature mapping methods can be divided into four main types (Fig. 3). First, decomposition-based methods decompose the feature data into new axes with lower redundancy or correlation. They include principal component analysis (102) and partial least square regression (103). These methods are usually very computationally efficient. Second, clustering-based methods group features into clusters according to their similarities, and then generate phenotypic profiles by combining features from the same clusters together (104,105). Third, transformation-based methods transform the data into new feature spaces while retaining certain inter-condition properties, such as the phenotypic dissimilarity among different tested conditions, in the original feature space. These methods include multi-dimensional scaling (106) and kernel-base methods (107). A more recently developed method is autoencoder, which is based on multilayer (or "deep") neural networks with a small central layer to reconstruct high-dimensional input data (108,109). Interestingly, autoencoder may also be applied directly to the raw pixel values in the images without the need to define or extract phenotypic features from the images. Finally, model-based methods

construct models to represent the observed phenotypes and use parameters of the models as phenotypic profiles. These models may represent the prevalence (in the form of histograms or distributions) of different cellular subpopulations (110,111) or local patches/key-point regions (89), temporal relationships between different classes of phenotypes (87,91), or cell shape or protein subcellular localization patterns (112). These methods have the advantage of being interpretable, and thus may be used to infer or identify the underlying mechanisms that generate the observed phenotypes.

Supervised Versus Unsupervised Profiling

After phenotypic profiling, the resulting profiles are usually used to construct computational models for the observed phenotypes (Fig. 1). In supervised learning, the models are trained using phenotypic profiles collected under intrinsic or extrinsic conditions that are known to induce certain pre-defined or categorized changes in cellular phenotypes (85), such as localization at specific subcellular compartments (88,89), stages of cell cycle (53), or responses to chemicals (40,113). Common computational methods for building supervised models include support vector machine (53,89) and random forest (40,114). Once trained, supervised models can be used to automatically assign new test samples into one of the pre-defined categories based on the measured phenotypic profiles. For example, supervised models were used to predict biological functions of new proteins (50,89), or pharmacological (113) and toxicological (40) effects of new chemicals. However, one of the major limitations of supervised models is that they require pre-defined phenotypic categories. As discussed earlier, the numbers and definitions of all possible categories of phenotypic changes may be unknown a priori.

To overcome this, unsupervised learning may be used to find phenotypic categories by grouping conditions with similar phenotypic profiles together. Commonly used methods for building unsupervised models include hierarchical or K-mean clustering (85). The identified categories may lead to the discovery of novel relationships among the tested conditions. For example, unsupervised models were used to discover proteins that are involved in related biological processes or functions, cellular states, common targets or mechanisms of chemicals (3,22,34,87). However, in practice, the occurrence of different phenotypic categories is usually non-uniform. Therefore, it remains a challenge to automatically identify "rare" phenotypic categories because insufficient numbers of measured conditions may be associated with these categories. Models based on skewed and heavy-tailed distributions (115) and spanning-tree progression analysis (116) have been proposed to identify "rare" clusters in flow cytometry data, and may also be used for clustering of image-based phenotypic profiles.

Another limitation of supervised learning is that large numbers of labeled samples are usually required to build generalizable and predictive models. The labeling of samples is usually a time- and labor-intensive process. To overcome this limitation, a semi-supervised learning method, called "active learning," may be used (117,118). The method interactively queries human annotators or performs additional perturbation experiments to obtain the labels of the most relevant samples, thereby achieving high prediction accuracy while minimizing the required number of labeled samples. For example, a phenotypic screening platform has recently been developed to integrate liquid handling robotics, automated microscopy, and active learning algorithms (118). The platform could accurately classify the effects of 48 chemical compounds by performing only 29% of all possible perturbation and imaging experiments (118). As phenotypic screens are designed to test increasing numbers of intrinsic and extrinsic factors, we expect to see more future adoptions of similar intelligent experimental design methods to "virtually" scale up the throughput of phenotypic screens.

Concluding Remarks

Several studies have found that phenotypic profiling can be used to discover novel phenotypes. For example, we have previously shown that P-profiles, which were constructed without using any pre-defined phenotypic category, can be used to search and rank proteins based on their dissimilarity in subcellular localization patterns to a set of query proteins (3) (Fig. 4). Other unsupervised studies have identified novel biologically active small molecules (34), temporal patterns during cell cycle (87), or subpopulations of cells with heterogeneous drug responses (110). Several previous studies have also compared the performances of supervised models based on phenotypic profiles or raw high-dimensional feature data. Phenotypic profiles were consistently found to give better classification performances in both intrinsic (3,89,90) or extrinsic (40,84) phenotype screens.

Despite the advantages of phenotypic profiling, it is not widely used in current large-scale image-based screens (32). One of the reasons is the lack of user-friendly software tools to perform the phenotypic profiling process. The OMERO (82), CellProfiler (83), and cellXpress (84) software packages can be used to generate and export large numbers of location-dependent and -independent features (Table 1), but customized computer programs or scripts are still needed to construct phenotypic profiles. Many of the profile construction methods are available as library packages under the R, Matlab, or python environments. The cellXpress software package also includes software routines for performing the D- and P-profiling algorithms under the R environment. Weka is also a powerful, general, and user-friendly software package that can be used to perform profile construction and data analysis (119). Another reason is that many different image features and methods have been used or developed for phenotypic profiling (see above sections). It is often unclear, especially for novice users, which image feature set or method is the most appropriate for a specific application. We expect that, as the field matures, more comparative studies will be performed and lead to more standardized feature sets and procedures for phenotypic profiling. Interestingly, several past studies have consistently found that Haralick's texture features are highly informative (22,40,86,88,91,97), and therefore these features

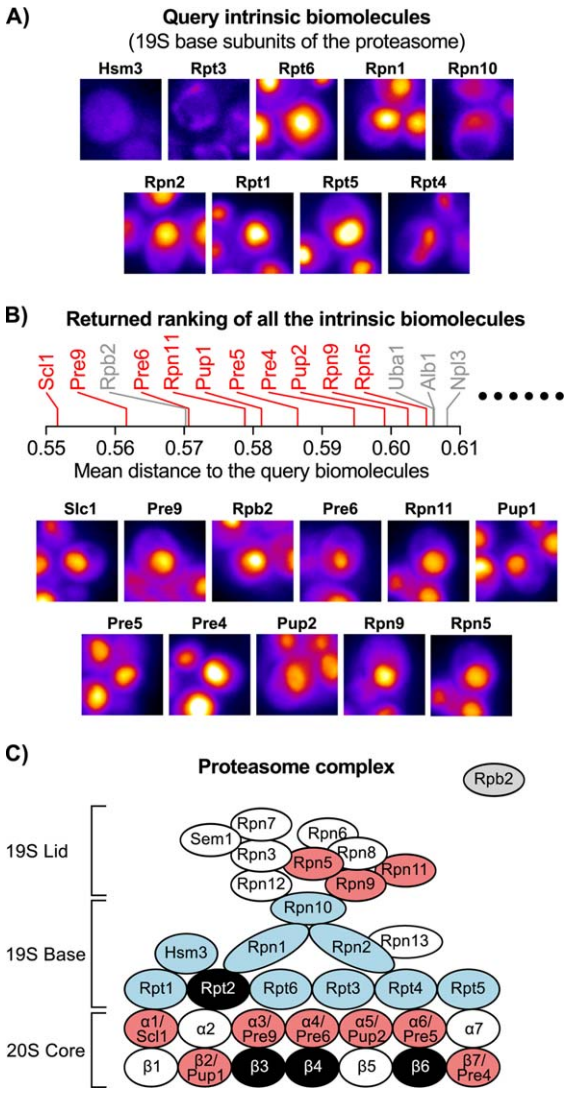

Figure 4. Phenotypic profiles of proteins can be used to identify intrinsic factors with similar subcellular localization patterns. An example output from the Protein Localization Analysis and Search Tool (PLAST), which is based on the P-profiling algorithm (3). The PLAST database (http://plast.bii.a-star.edu.sg) consists of the images of ~4,000 GFP-tagged budding yeast proteins (2), and was built without any pre-defined category of subcellular localization patterns. (A) The inputs (or query genes) to the PLAST database are the gene symbols for nine genes that are known to constitute the 19S base subunit of the proteasome complex. Shown are the images of these query genes. (B) The output of the search is a ranking of all the ~4,000 genes (except for the query genes) according to their average distances to the query genes computed based on the P-profiles of the genes. Shown are the distances (top) and images (bottom) of the top hits returned from PLAST. (C) Shown is a schematic of all the known subunits of the proteasome complex, which are arranged approximately according to their known physical interactions or structures within the complex. Most of the top hits returned from PLAST are from the 19S lid and 20S core sub-complexes. Blue = query 19S base subunits, red = detected known subunits of proteasome, white- = missed known subunits, gray = other detected proteins, black- = known subunits without PLAST profiles. PLAST was found to be significantly better than existing protein-localization annotations in recovering the subunits of ~33% of the 197 known protein complexes in the budding yeast (3). [Color figure can be viewed at wileyonlinelibrary.com]

are recommended to be included as part of the standard initial feature set in phenotypic profiling. Finally, image-based features, especially location-dependent features, may not be easily interpretable. As mentioned above, this problem may be addressed using model-based phenotypic profiles. Nevertheless, as with all other types of screening methodologies, further biomolecular validation experiments are still required to verify the identified hits and understand the underlying mechanisms that generate the phenotypes.

Literature Cited

1. Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, et al. High-dimensional and large-scale phenotyping of yeast mutants. Proc Natl Acad Sci U S A 2005;102:19015–19020.
2. Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. Global analysis of protein localization in budding yeast. Nature 2003;425:686–691.
3. Loo L-H, Laksameethanasan D, Tung Y-L. Quantitative protein localization signatures reveal an association between spatial and functional divergences of proteins. PLoS Comput Biol 2014;10:e1003504.
4. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 2003;21:697–700.
5. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. Nature 2006;441:840–846.
6. Giepmans BNG, Adams SR, Ellisman MH, Tsien RY. The fluorescent toolbox for assessing protein location and function. Science 2006;312:217–224.
7. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 2010;28:1248–1250.
8. Buszczak M, Paterno S, Lighthouse D, Bachman J, Planck J, Owen S, Skora AD, Nystul TG, Ohlstein B, Allen A, et al. The Carnegie protein trap library: A versatile tool for drosophila developmental studies. Genetics 2007;175:1505–1531.
9. Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, Sigal A, Milo R, Cohen-Saidon C, Liron Y, Kam Z, et al. Dynamic proteomics of individual cancer cells in response to a drug. Science 2008;322:1511–1516.
10. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. A sampling of the yeast proteome. Mol Cell Biol 1999;19:7357–7368.
11. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 2007;25:117–124.
12. Levsky JM, Singer RH. Fluorescence in situ hybridization: past, present and future. J Cell Sci 2003;116:2833–2838.
13. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods 2008;5:877–879.
14. Femino AM, Fay FS, Fogarty K, Singer RH. Visualization of single RNA transcripts in situ. Science 1998;280:585–590.
15. Battich N, Stoeger T, Pelkmans L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nat Methods 2013;10:1127–1133.
16. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015;348:aaa6090.
17. Cajigas IJ, Tushev G, Will TJ, tom Dieck S, Fuerst N, Schuman EM. The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging. Neuron 2012;74:453–466.
18. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, Krause HM. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell 2007;131:174–187.
19. Shankavaram UT, Reinhold WC, Nishizuka S, Major S, Morita D, Chary KK, Reimers MA, Scherf U, Kahn A, Dolginow D, et al. Transcript and protein expression profiles of the NCI-60 cancer cell panel: An integromic microarray study. Mol Cancer Ther 2007;6:820–832.
20. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature 2011;473:337–342.
21. Niepel M, Hafner M, Pace EA, Chung M, Chai DH, Zhou L, Schoeberl B, Sorger PK. Profiles of Basal and stimulated receptor signaling networks predict drug response in breast cancer lines. Sci Signal 2013;6:ra84
22. Loo L-H, Wu LF, Altschuler SJ. Image-based multivariate profiling of drug responses from single cells. Nat Methods 2007;4:445–453.
23. Simpson JC, Joggerst B, Laketa V, Verissimo F, Cetin C, Erfle H, Bexiga MG, Singan VR, Hériché J-K, Neumann B, et al. Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. Nat Cell Biol 2012;14:764–774.
24. Paige JS, Wu KY, Jaffrey SR. RNA mimics of green fluorescent protein. Science 2011;333:642–646.
25. Liu HS, Jan MS, Chou CK, Chen PH, Ke NJ. Is green fluorescent protein toxic to the living cells? Biochem Biophys Res Commun 1999;260:712–717.
26. Ratz M, Testa I, Hell SW, Jakobs S. CRISPR/Cas9-mediated endogenous protein tagging for RESOLFT super-resolution microscopy of living human cells. Sci Rep 2015;5:9592.

27. Nelles DA, Fang MY, O'Connell MR, Xu JL, Markmiller SJ, Doudna JA, Yeo GW. Programmable RNA tracking in live cells with CRISPR/Cas9. Cell 2016;165:488–496.
28. Lukinavičius G, Reymond L, D'Este E, Masharina A, Göttfert F, Ta H, Güther A, Fournier M, Rizzo S, Waldmann H, et al. Fluorogenic probes for live-cell imaging of the cytoskeleton. Nat Methods 2014;11:731–733.
29. Stockwell BR. Chemical genetics: Ligand-based discovery of gene function. Nat Rev Genet 2000;1:116–125.
30. Drews J. Drug discovery: A historical perspective. Science 2000;287:1960–1964.
31. Swinney DC, Anthony J. How were new medicines discovered? Nat Rev Drug Discov 2011;10:507–519.
32. Singh S, Carpenter AE, Genovesio A. Increasing the content of high-content screening: An overview. J Biomol Screen 2014;19:640–650.
33. Link W, Oyarzabal J, Serelde BG, Albarran MI, Rabal O, Cebriá A, Alfonso P, Fominaya J, Renner O, Peregrina S, et al. Chemical interrogation of FOXO3a nuclear translocation identifies potent and selective inhibitors of phosphoinositide 3-kinases. J Biol Chem 2009;284:28392–28400.
34. Tanaka M, Bateman R, Rauh D, Vaisberg E, Ramachandani S, Zhang C, Hansen KC, Burlingame AL, Trautman JK, Shokat KM, et al. An unbiased cell morphology-based screen for new, biologically active small molecules. PLoS Biol 2005;3:e128.
35. Mudhasani R, Kota KP, Retterer C, Tran JP, Whitehouse CA, Bavari S. High content image-based screening of a protease inhibitor library reveals compounds broadly active against rift valley fever virus and other highly pathogenic RNA viruses. PLoS Negl Trop Dis 2014;8:e3095.
36. Schulz MMP, Reisen F, Zgraggen S, Fischer S, Yuen D, Kang GJ, Chen L, Schneider G, Detmar M. Phenotype-based high-content chemical library screening identifies statins as inhibitors of in vivo lymphangiogenesis. Proc Natl Acad Sci 2012;109: E2665–E2674.
37. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. Science 2004;306:1194–1198.
38. Kleinstreuer NC, Yang J, Berg EL, Knudsen TB, Richard AM, Martin MT, Reif DM, Judson RS, Polokoff M, Dix DJ, et al. Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. Nat Biotechnol 2014;32:583–591.
39. Cleuvers M. Mixture toxicity of the anti-inflammatory drugs diclofenac, ibuprofen, naproxen, and acetylsalicylic acid. Ecotoxicol Environ Saf 2004;59:309–315.
40. Su R, Xiong S, Zink D, Loo L-H. High-throughput imaging-based nephrotoxicity prediction for xenobiotics with diverse chemical structures. Arch Toxicol 2015;1–16.
41. Hamilton AJ, Baulcombe DC. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science 1999;286:950–952.
42. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. Science 2007;315:1709–1712.
43. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 2012;337:816–821.
44. Kawasaki H, Taira K. Induction of DNA methylation and gene silencing by short interfering RNAs in human cells. Nature 2004;431:211–217.
45. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell 2014;159:647–661.
46. Friedman A, Perrimon N. A functional RNAi screen for regulators of receptor tyrosine kinase and ERK signalling. Nature 2006;444:230–234.
47. Green RA, Kao H-L, Audhya A, Arur S, Mayers JR, Fridolfsson HN, Schulman M, Schloissnig S, Niessen S, Laband K, et al. A high-resolution *C. elegans* essential gene network based on phenotypic profiling of a complex tissue. Cell 2011;145:470–482.
48. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 2014;343:84–87.
49. Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, Balija V, O'Shaughnessy A, Gnoj L, Scobie K, et al. A resource for large-scale RNA-interference-based screens in mammals. Nature 2004;428:427–431.
50. Neumann B, Walter T, Hériché J-K, Bulkescher J, Erfle H, Conrad C, Rogers P, Poser I, Held M, Liebel U, et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature 2010;464:721–727.
51. Simpson KJ, Selfors LM, Bui J, Reynolds A, Leake D, Khvorova A, Brugge JS. Identification of genes that regulate epithelial cell migration using an siRNA screening approach. Nat Cell Biol 2008;10:1027–1038.
52. Pelkmans L, Fava E, Grabner H, Hannus M, Habermann B, Krausz E, Zerial M. Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. Nature 2005;436:78–86.
53. Neumann B, Held M, Liebel U, Erfle H, Rogers P, Pepperkok R, Ellenberg J. High-throughput RNAi screening by time-lapse imaging of live human cells. Nat Methods 2006;3:385–390.
54. Morgens DW, Deans RM, Li A, Bassik MC. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. Nat Biotechnol 2016;34:634–636.
55. Raab RM, Stephanopoulos G. Dynamics of gene silencing by RNA interference. Biotechnol Bioeng 2004;88:121–132.
56. Jacobsen L, Calvin S, Lobenhofer E. Transcriptional effects of transfection: The potential for misinterpretation of gene expression data generated from transiently transfected cells. BioTechniques 2009;47:617–624.
57. Knight ZA, Lin H, Shokat KM. Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 2010;10:130–137.
58. Conrad C, Gerlich DW. Automated microscopy for high-content RNAi screening. J Cell Biol 2010;188:453–461.
59. Michael S, Auld D, Klumpp C, Jadhav A, Zheng W, Thorne N, Austin CP, Inglese J, Simeonov A. A robotic platform for quantitative high-throughput screening. Assay Drug Dev Technol 2008;6:637–657.
60. Shen F, Hodgson L, Hahn K. Digital autofocus methods for automated microscopy. Methods Enzymol 2006;414:620–632.
61. Liron Y, Paran Y, Zatorsky NG, Geiger B, Kam Z. Laser autofocusing system for high-resolution cell biological imaging. J Microsc 2006;221:145–151.
62. Conrad C, Wünsche A, Tan TH, Bulkescher J, Sieckmann F, Verissimo F, Edelstein A, Walter T, Liebel U, Pepperkok R, et al. Micropilot: Automation of fluorescence microscopy-based imaging for systems biology. Nat Methods 2011;8:246–249.
63. Mathews LA, Keller JM, Goodwin BL, Guha R, Shinn P, Mull R, Thomas CJ, Kluyver RL, de Sayers TJ, Ferrer M. A 1536-well quantitative high-throughput screen to identify compounds targeting cancer stem cells. J Biomol Screen 2012;17:1231–1242.
64. Dittrich PS, Manz A. Lab-on-a-chip: Microfluidics in drug discovery. Nat Rev Drug Discov 2006;5:210–218.
65. Hong J, Edel JB, deMello AJ. Micro- and nanofluidic systems for high-throughput biological screening. Drug Discov Today 2009;14:134–146.
66. Waters JC. Accuracy and precision in quantitative fluorescence microscopy. J Cell Biol 2009;185:1135–1148.
67. Preibisch S, Saalfeld S, Tomancak P. Globally optimal stitching of tiled 3D microscopic image acquisitions. Bioinformatics 2009;25:1463–1465.
68. Meijering E. Cell segmentation: 50 years down the road [life sciences]. IEEE Signal Process Mag 2012;29:140–145.
69. Yu W, Lee HK, Hariharan S, Bu W, Ahmed S. Evolving generalized Voronoi diagrams for accurate cellular image segmentation. Cytom Part J Int Soc Anal Cytol 2010;77: 379–386.
70. Yu W, Lee HK, Hariharan S, Bu W, Ahmed S. Quantitative neurite outgrowth measurement based on image segmentation with topological dependence. Cytometry Part A 2009;75A:289–297.
71. Yu W, Lee HK, Hariharan S, Bu W, Ahmed S. Level set segmentation of cellular images based on topological dependence. In: Bebis G, Boyle R, Parvin B, Koracin D, Remagnino P, Porikli F, Peters J, Klosowski J, Arns L, Chun YK, Rhyne T-M, Monroe L, editors. Advances in Visual Computing, Lecture Notes in Computer Science. Berlin-Heidelberg: Springer; 2008. pp 540–551.
72. Law YN, Lee HK, Liu C, Yip AM. A variational model for segmentation of overlapping objects with additive intensity value. IEEE Trans Image Process 2011;20:1495–1503.
73. Law YN, Yip AM, Lee HK. Automatic measurement of volume percentage stroma in endometrial images using texture segmentation. J Microsc 2011;241:171–178.
74. Law YN, Ogg S, Common J, Tan D, Lane EB, Yip AM, Lee HK. Automated protein distribution detection in high-throughput image-based siRNA library screens. J Signal Process Syst 2008;55:1–13.
75. Ledley RS. High-speed automatic analysis of biomedical pictures. Science 1964;146: 216–223.
76. Belson M, Dudley AW Jr, Ledley RS. Automatic computer measurements of neurons. Pattern Recognit 1968;1:119–128.
77. Andrade R, Crisol L, Prado R, Boyano MD, Arluzea J, Aréchaga J. Plasma membrane and nuclear envelope integrity during the blebbing stage of apoptosis: A time-lapse study. Biol Cell Auspices Eur Cell Biol Organ 2010;102:25–35.
78. Matassov D, Kagan T, Leblanc J, Sikorska M, Zakeri Z. Measurement of apoptosis by DNA fragmentation. Methods Mol Biol Clifton NJ 2004;282:1–17.
79. Blaecke A, Delneste Y, Herbault N, Jeannin P, Bonnefoy J-Y, Beck A, Aubry J-P. Measurement of nuclear factor-kappa B translocation on lipopolysaccharide-activated human dendritic cells by confocal microscopy and flow cytometry. Cytometry 2002; 48:71–79.
80. Giuliano KA, DeBiasio RL, Dunlay RT, Gough A, Volosky JM, Zock J, Pavlakis GN, Taylor DL. High-content screening: A new approach to easing key bottlenecks in the drug discovery process. J Biomol Screen 1997;2:249–259.
81. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods 2012;9:671–675.
82. Allan C, Burel J-M, Moore J, Blackburn C, Linkert M, Loynton S, MacDonald D, Moore WJ, Neves C, Patterson A, et al. OMERO: Flexible, model-driven data management for experimental biology. Nat Methods 2012;9:245–253.
83. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM. CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. Genome Biol 2006;7: R100.
84. Laksameethanasan D, Tan RZ, Toh GW, Loo L-H. cellXpress: A fast and user-friendly software platform for profiling cellular phenotypes. BMC Bioinf 2013;14:S4.
85. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S. Machine learning and its applications to biology. PLoS Comput Biol 2007;3:e116.
86. Fuchs F, Pau G, Kranz D, Sklyar O, Budjan C, Steinbrink S, Horn T, Pedal A, Huber W, Boutros M. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. Mol Syst Biol 2010;6:370.
87. Zhong Q, Busetto AG, Fededa JP, Buhmann JM, Gerlich DW. Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. Nat Methods 2012;9: 711–713.
88. Boland MV, Murphy RF. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics 2001;17:1213–1223.
89. Coelho LP, Kangas JD, Naik AW, Osuna-Highley E, Glory-Afshar E, Fuhrman M, Simha R, Berget PB, Jarvik JW, Murphy RF. Determining the subcellular location of new proteins from microscope images using local features. Bioinformatics 2013;29: 2343–2349.

90. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;SMC-3:610–621.
91. Held M, Schmitz MHA, Fischer B, Walter T, Neumann B, Olma MH, Peter M, Ellenberg J, Gerlich DW. CellCognition: Time-resolved phenotype annotation in high-throughput live cell imaging. Nat Methods 2010;7:747–754.
92. Manjunath BS, Ma WY. Texture features for browsing and retrieval of image data. IEEE Trans Pattern Anal Mach Intell 1996;18:837–842.
93. Khotanzad A, Hong YH. Invariant image recognition by Zernike moments. IEEE Trans Pattern Anal Mach Intell 1990;12:489–497.
94. Shamir L, Orlov N, Eckley DM, Macura T, Johnston J, Goldberg IG. Wndchrm—An open source utility for biological image analysis. Source Code Biol. Med 2008;3:13.
95. Lowe DG. Object recognition from local scale-invariant features. In: Proceedings of Seventh International Conference on Computer Vision, Kerkyra, Greece. Vol 2. IEEE; 1999. pp 1150–1157.
96. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–1182.
97. Huang K, Velliste M, Murphy RF. Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. Proc SPIE 2003; 4962:307–318.
98. Kira K, Rendell LA. A practical approach to feature selection. In: Proceedings of Ninth International Workshop on Machine Learning ML92. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1992. pp 249–256.
99. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.
100. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognit Lett 1994;15:1119–1125.
101. Siedlecki W, Sklansky J. A note on genetic algorithms for large-scale feature selection. Pattern Recognit Lett 1989;10:335–347.
102. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol 1933;24:417–441.
103. Boulesteix A-L, Strimmer K. Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 2007;8:32–44.
104. Hunter LA, Krafft S, Stingo F, Choi H, Martel MK, Kry SF, Court LE. High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. Med Phys 2013;40:121916.
105. Parmar C, Leijenaar RTH, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, Rietbergen MM, Haibe-Kains B, Lambin P, Aerts HJWL. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Sci Rep 2015;5: 11044.
106. Torgerson WS. Multidimensional scaling: I. Theory and method. Psychometrika 1952;17:401–419.
107. Schölkopf B, Smola A, Müller K-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Comput 1998;10:1299–1319.
108. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313:504–507.
109. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. In: Proceedings of 25th International Conference on Machine Learning ICML'08. New York, NY: ACM; 2008. pp 1096–1103.
110. Slack MD, Martinez ED, Wu LF, Altschuler SJ. Characterizing heterogeneous cellular responses to perturbations. Proc Natl Acad Sci U S A 2008;105:19306–19311.
111. Snijder B, Sacher R, Rämö P, Damm E-M, Liberali P, Pelkmans L. Population context determines cell-to-cell variability in endocytosis and virus infection. Nature 2009;461:520–523.
112. Peng T, Murphy RF. Image-derived, three-dimensional generative models of cellular organization. Cytometry Part A 2011;79A:383–391.
113. Kang J, Hsu C-H, Wu Q, Liu S, Coster AD, Posner BA, Altschuler SJ, Wu LF. Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines. Nat Biotechnol 2016;34:70–77.
114. Breiman L. Random forests. Mach. Learn 2001;45:5–32.
115. Pyne S, Hu X, Wang K, Rossin E, Lin T-I, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, et al. Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci U S A 2009;106:8519–8524.
116. Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol 2011;29:886–891.
117. Warmuth MK, Liao J, Rätsch G, Mathieson M, Putta S, Lemmen C. Active learning with support vector machines in the drug discovery process. J Chem Inf Comput Sci 2003;43:667–673.
118. Naik AW, Kangas JD, Sullivan DP, Murphy RF. Active machine learning-driven experimentation to determine compound effects on protein patterns. eLife 2016;5: e10047.
119. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. SIGKDD Explor Newsl 2009;11:10–18.