

Molecular Cancer Therapeutics



High-Content Phenotypic Profiling of Drug Response Signatures across Distinct Cancer Cells

Peter D. Caie, Rebecca E. Walls, Alexandra Ingleston-Orme, et al.

Mol Cancer Ther 2010;9:1913-1926. Published OnlineFirst June 8, 2010.

Updated Version	Access the most recent version of this article at: doi: 10.1158/1535-7163.MCT-09-1148
Supplementary Material	Access the most recent supplemental material at: http://mct.aacrjournals.org/content/suppl/2010/06/07/1535-7163.MCT-09-1148.DC1.html

Cited Articles	This article cites 28 articles, 5 of which you can access for free at: http://mct.aacrjournals.org/content/9/6/1913.full.html#ref-list-1
-----------------------	---

E-mail alerts	Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions	To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org .
Permissions	To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org .

High-Content Phenotypic Profiling of Drug Response Signatures across Distinct Cancer Cells

Peter D. Caie¹, Rebecca E. Walls², Alexandra Ingleston-Orme¹, Sandeep Daya¹, Tom Houslay¹, Rob Eagle³, Mark E. Roberts⁴, and Neil O. Carragher¹

Abstract

The application of high-content imaging in conjunction with multivariate clustering techniques has recently shown value in the confirmation of cellular activity and further characterization of drug mode of action following pharmacologic perturbation. However, such practical examples of phenotypic profiling of drug response published to date have largely been restricted to cell lines and phenotypic response markers that are amenable to basic cellular imaging. As such, these approaches preclude the analysis of both complex heterogeneous phenotypic responses and subtle changes in cell morphology across physiologically relevant cell panels. Here, we describe the application of a cell-based assay and custom designed image analysis algorithms designed to monitor morphologic phenotypic response in detail across distinct cancer cell types. We further describe the integration of these methods with automated data analysis workflows incorporating principal component analysis, Kohonen neural networking, and kNN classification to enable rapid and robust interrogation of such data sets. We show the utility of these approaches by providing novel insight into pharmacologic response across four cancer cell types, Ovcar3, MiaPaCa2, and MCF7 cells wild-type and mutant for p53. These methods have the potential to drive the development of a new generation of novel therapeutic classes encompassing pharmacologic compositions or polypharmacology in appropriate disease context. *Mol Cancer Ther*; 9(6); 1913–26. ©2010 AACR.

Introduction

High throughput target-directed drug discovery has been the research paradigm favored by the pharmaceutical and biotechnology sectors over the last two decades. Despite the increased identification of putative therapeutic targets in the post-genomics era, widespread adoption of target-directed drug discovery has been accompanied by a steady decline in the approval of drugs against new targets and a significant increase in the attrition of candidate drugs during preclinical and clinical development as a result of toxicity and poor efficacy response (1, 2). Target-directed screening has encouraged the application of simplistic biochemical or engineered cellular assays. Although these approaches are amenable to high-throughput screening and have resulted in the discovery

of potent and highly selective agonists and antagonists, they provide limited information on how therapeutics influence complex physiologic systems. Such limitations are a contributing factor to high attrition rates at later stages in the drug development process.

Advances in automated high-content imaging and image analysis methods offer an alternative to the traditional target-directed screening approach (3–5). Application of imaging technologies to profile the phenotypic response to molecular or pharmacologic perturbation at cellular or tissue levels enables the study of therapeutic response irrespective of putative target activity (6, 7). Recent studies have used high-content cellular assays to show the value of phenotypic profiling to first confirm cellular activity and second, in conjunction with multivariate clustering techniques, to elucidate mechanism of action (MOA) in which drug mode of action or primary targets are unknown (8–10).

Acquired drug resistance and intrinsic heterogeneity among patients, tumor sites, and stage of disease represent significant challenges to the successful treatment of cancer with targeted pharmacologic agents (11–13). To advance the discovery and development of effective anti-tumor agents, it will be necessary to recapitulate tumor heterogeneity into early drug discovery. The phenotypic analysis of pharmacologic response across panels of distinct cancer cell lines that act as surrogates for heterogeneous patient populations or distinct cancer phenotypes may assist the development of therapeutic or pharmaceutical

Authors' Affiliations: ¹Advanced Science and Technology Laboratory and ²Discovery Statistics, AstraZeneca R&D Charnwood, Loughborough, United Kingdom; ³GlaxoSmithKline, Molecular Discovery Research, Harlow, Essex, United Kingdom; and ⁴Tessella plc, 26 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, United Kingdom

Note: Supplementary material for this article is available at Molecular Cancer Therapeutics Online (<http://mct.aacrjournals.org/>).

Corresponding Author: Neil Carragher, Advanced Science and Technology Laboratory, AstraZeneca R&D Charnwood, Loughborough, LE11 5RH England, UK. Phone: 440-1509645408; Fax: 440-1509647709. E-mail: neil.carragher@astrazeneca.com

doi: 10.1158/1535-7163.MCT-09-1148

©2010 American Association for Cancer Research.

compositions targeting redundant tumor resistance mechanisms or distinct genetic and epigenetic traits (14–16). To date practical examples of high-content phenotypic profiling have largely been restricted to cell lines and phenotypic response markers, including exogenously expressed proteins, which are amenable to predefined image analysis solutions (8, 17, 18). Most high-content assays and associated image analysis algorithms are manually optimized for a specific cell line or use a limited parameter set restricted to generic features independent of distinct cellular morphology. These approaches preclude the detailed analysis of complex heterogeneous responses or subtle changes in cell morphology induced by selective pharmacologic agents across physiologically relevant cell panels.

In this report, we describe the development and application of a multiparametric high-content cell-based assay and bespoke custom-designed, image analysis algorithms to enable detailed monitoring of changes to cytoskeletal, nuclear, and cellular morphologies across heterogeneous cancer cell types. A significant bottleneck in the analysis of multiparametric high-content phenotypic assays is data processing, including quality validation, normalization, and secondary multivariate statistical analysis (19). We describe the application of three novel multiparametric high-content analysis methods integrated into an automated data analysis workflow: Mahalanobis hit stratification to identify active compound and doses as determined by the multiparametric phenotypic response measurements, Kohonen neural network analysis to monitor multiparametric phenotypic response across dose response and cell types, and KNN classification to rank compound similarity and predict MOA as determined by a multiparametric image analysis readout. These methods enable the rapid interrogation of complex high-content phenotypic response data across multiple cell lines within the time frame required by the drug-screening cascade. We show the utility of these approaches by profiling the phenotypic response of four distinct cancer cell lines following perturbation with a small-molecule compound library. Our data and analysis methods provide novel insight into pharmacologic response across the following four cancer cell lines representing distinct tissue origin and p53 mutation status: Ovar3 (ovarian), MiaPaCa2 (pancreatic), and MCF7 cells wild-type and mutant for p53 (breast).

Materials and Methods

Cell culture

The cell lines MCF7-wt (breast cancer expressing wild-type p53), MCF7-p53 (transfected dominant-negative truncated p53 mutant gene), MiaPaCa2 (pancreatic cancer), and Ovar3 (ovarian cancer) were subcultured in the following media preparations: MCF7-wt and MCF7-p53 with RPMI 1640, 10% fetal bovine serum, 1% GlutaMAX (200 mmol/L), and 900 µg/mL G418;

MiaPaCa2 with DMEM, 10% fetal bovine serum, and 1% GlutaMAX (200 mmol/L); and Ovar3 with RPMI 1640, 10% fetal bovine serum, and 1% GlutaMAX (200 mmol/L). All media and supplements were supplied by Sigma. All cells were maintained at 37°C, 5% CO₂, and 100% humidity. For high-content assays, cells were resuspended in Accutase (Sigma) and plated on collagen-coated black 96-well clear bottom plates (Becton Dickinson) using a WellMate-fine bore (Matrix) at 5,000 cells per well in 95 µL media, with the exception of Ovar3, which were plated at 12,000 cells per well. These plates were subsequently cultured at 37°C, 5% CO₂, and 100% humidity for 24 hours before compound addition. All cell lines were originally obtained from the American Type Culture Collection. MCF7-wt cells were genetically engineered to express a dominant-negative, transactivation-deficient p53 mutant and stable expression clones were derived and subcultured in antibiotic (900 µg/mL G418) selection media. All cell lines were subsequently authenticated as representing original parental derivatives through barcode sequencing of conserved DNA segments and comparison with the American Type Culture Collection DNA profiles. All cell line DNA sequencing was done during 2009 by LGC Ltd. using standardized ABI3730xl sequencing platform.

Compound treatment

All compounds were prepared and titrated as an eight-point half-log dose response in a 96-well microtiter plate in 100% DMSO at 1,000× final drug concentration. Highest concentrations for each compound were selected from previously published data indicating activity in cell-based assays. Compound dilutions were prepared using a Tecan-automated dispensing system. Using a Biomek Fx-automated dispensing system, 5 µL from the compound dilution plate were dispensed into the intermediate plate containing 250 µL of cell culture media and mixed before transfer of 5 µL to each cell plate. Cell plates were incubated with the compound and 0.1% DMSO control at 37°C, 5% CO₂, and 100% humidity for 24 and 48 hours.

Kinetic imaging

Kinetic profiling of compound responses across cell types was done using the Cell-IQ instrument (Chip-Man Technologies Ltd). The Cell-IQ represents a fully integrated incubator, phase-contrast image acquisition, and artificial intelligent solution for kinetic profiling phenotypic response following chemical perturbation. Cells were seeded at 7.5 × 10³ cells per well in 48-multiwell plates (Nunclon) for 24 hours before compound addition. Three regions of interest per well were then imaged by the Cell-IQ every 20 minutes over a 120-hour period. Supplementary data presented represents a consistent example from one field of view for each compound treatment.

Immunocytochemistry

All immunostaining procedures were done at room temperature in 96-well plates. All volumes are 100 µL

unless otherwise stated. Cells were fixed following direct addition of 100 μ L 8% paraformaldehyde in PBS to cells (final concentration, 4%) and were incubated for 20 minutes. Cells were washed three times in PBS and incubated in blocking buffer (PBS containing 1.1% bovine serum albumin, and 0.2% Triton X100) for 30 minutes. Primary antibody solution diluted in blocking buffer (1:500; mouse anti- β -tubulin IgG1, Sigma T5293) was added at 40 μ L per well and incubated for 1 hour. Cells were washed three times with blocking buffer and incubated with blocking buffer for 30 minutes. Secondary antibody solution of (40 μ L) Alexa Fluor 488 donkey anti-mouse IgG (H+L; Molecular probes A21202; diluted at 1:500), Phalloidin conjugated to Alexa Fluor 568 (Molecular probes A12380; diluted 1:500), and 4',6-diamidino-2-phenylindole (DAPI; Sigma D8417) diluted at 1:250 in blocking buffer was incubated with cells for 45 minutes in the dark. Cells were washed three times with PBS. Deep red HCS cell mask (Molecular probes, H34560) diluted at 1:120,000 was added and incubated for 10 minutes. Cells were washed twice with PBS. Plates were sealed with black plate sealers before imaging.

Fluorescent image acquisition and image analysis

All images were acquired on the automated ImageXpress 5000A high-content imaging platform (Molecular Devices) using a $\times 20$ PanFluor ELWD Ph1 DM objective and a 16-bit camera binning resolution of 1. Four separate fields of view per well were acquired using laser-based autofocus parameters optimized for cell plates and cell type. Image analysis was done using custom-designed analysis algorithms created using Definiens Cellenger cognition network software. Images were processed for analysis as follows.

Step 1: Quality control check. Using expected values for both mean and SD of intensities for the DAPI and actin layers and the mean intensity of the cell mask layer, an aggregated score was created. This score must be above a designated threshold for the image to be of sufficient quality to proceed. This process was run automatically to ensure that each image was of acceptable quality to enable accurate cell segmentation.

Step 2: Nuclear segmentation. A combination of pixel contrast and intensity values in the DAPI channel is used to find nuclear area. Reshaping methods based on iterative setting of shape and intensity metrics on the DAPI channel is implemented to separate discrete nuclei and accurately segment across a range of diverse nuclear phenotypes from control and compound perturbed cell lines.

Step 3: Cell assignment and cytoplasmic segmentation. Definiens works on an object-based method, and so a higher object level, based on cell mask and phalloidin-associated parameters is created above the "nuclei level" to allocate nuclei to their respective cells using the distance watershed technique. Cell segmentation was achieved by defining regions of interest above the background and

calculating a log score of the product of the three non-DAPI channels and disregarding the areas below a defined threshold.

Step 4: Subpopulation classification. Cells are subclassified as "other" or "round," depending on brightness levels and shape metrics. A manual shape metric threshold was set to ensure that other cells were not elliptical or circular in morphology. A customized feature within Definiens is used to determine cells that have an elongated and irregular cytoskeletal morphology, through dividing the "shape index" of an object by its "density." Shape index is defined within Definiens as an object's border length divided by four times the square root of its area (indicating "smoothness of border") and "density" is defined as the number of pixels forming the image object divided by its approximated radius based on the covariance matrix ("spatial distribution of the pixels of an image object"). It follows that for cells containing "pseudopods," there should be a lower density due to the larger spatial spread of pixels and a higher shape index as the border should be more ragged due to the presence of pseudopod protrusions. Dividing shape index by density should therefore give a higher figure for such cells. Shape index and density are subsequently determined for both cell mask and phalloidin staining, which enables a more robust detection of lamellipodia and pseudopodia structures. If there is a significant difference in area and shape/density metric between cell mask and phalloidin layers (reflecting more elongated and irregular cytoskeletal morphology), the cell is classified as having pseudopods; if there is no significant difference, then a "normal" phenotype is assigned. Visual inspection of this automated classification operation confirms that cells automatically classified as pseudopod represent elongated cells containing significantly more pseudopod extensions than normal counterparts.

Step 5: Individual cells were segmented and 150 distinct parameters, including DAPI nuclear intensity, nuclear area, Phalloidin intensity cytoplasm/nucleus, tubulin intensity cytoplasm/nucleus, cell circumference, cell number, etc., were calculated. In addition, we calculated round, normal, and pseudopod sub classifications as proportions of total cells. See Supplementary Fig. S2 for the full list of measured parameters.

Statistical and neural network analysis

Data aggregation, normalization, principal components analysis, and Mahalanobis hit stratification.

Step 1: All image analysis measurements were averaged over the cell population from each individual microwell sample to give "well-level" measurements.

Step 2: To account for week-to-week variability in intensity-based measurements, data were normalized, on a feature-by-feature basis, by subtracting the median of negative controls for each microtiter plate.

Step 3: Triplicate samples across separate microtiter plates were averaged to produce median values for each measured parameter.

Step 4: As an initial tool for exploring and visualizing multiparametric image analysis data, principal component analysis (PCA) was run in R and applied to all available data, including all compound and control DMSO-treated samples, on a per cell line and per time point basis.

Step 5: Outliers to the negative control cloud (DMSO control samples) were identified from the principal components plot and the corresponding well image examined manually. If the image suggested experimental artifact, the observation was removed.

Step 6: For each individual cell type and time point, a Mahalanobis distance metric (20) was applied across all compound and control DMSO-treated samples as a multiparametric hit stratification tool. We define activity in our phenotypic assay as something that is significantly different from the cloud of DMSO-negative controls. We calculate the Mahalanobis distance in multidimensional space between every negative control observation and the center of the DMSO cloud. We then calculate the Mahalanobis distance between the sample observation and the center of the DMSO control cloud. If the observation is very similar to the negative controls, this distance will be small. Alternatively, if the compound is doing something significantly different from DMSO control, the distance will be large. Repeating this process for every compound at each of its doses yields a vector of distances. A Mahalanobis distance threshold can then be applied to identify atypical responses either in shape or distance from the DMSO control. A manually defined Mahalanobis metric threshold at a significance value of $P < 0.01$ was used as the basis of an automatic hit stratification tool to identify any significant difference between compound-treated samples from the DMSO control cloud for each cell type and time point. A compound was deemed active if the Mahalanobis distance was significantly greater than the selected threshold for at least one of its doses tested.

Kohonen network. Multiparametric phenotypic response data across all cell types and time points were used to generate a two-dimensional Kohonen network (or self-organizing map) to facilitate the comparative analysis of phenotypic drug response across cell types, time points, and compound dose. A Kohonen network was created using the Spotfire Decision-Site software.

Step 1: The data used for the Kohonen network training were taken from the point after the multiple fields of view had been averaged and after the normalization of intraplate effects.

Step 2: Z-score normalization (subtracting the mean and then dividing by the SD) was then applied to each of the 150 image features before training. This ensured that features with higher magnitudes did not dominate the weighted-sum calculations used by the training algorithm.

Step 3: The map size was set to a 50×50 grid and a neighborhood size of 40 was used.

Step 4: A Gaussian neighborhood function was used with an initial radius size of 30, decreasing over the iter-

tions to a radius of 2. This linearly decreasing radius function was used to allow high-level structure to emerge more quickly than using a constant radius size.

Step 5: The learning rate was set at 0.05.

Step 6: The network was allowed to train for 50,000 iterations although it showed no significant further convergence after 40,000 iterations.

Step 7: The resultant map was used to visualize the multiparametric data points for each compound dose.

k-nearest neighbor classification. Finally, a *k*-nearest neighbor classification algorithm (kNN) was applied to the predictor variables on their original scale to make a prediction about a particular compound's MOA. For an object (compound) with an unknown classification, we first determine which of the training set are closest (most similar) to the new object and then assign the new object to the class that is most common among its *k* nearest neighbors. The algorithm was applied on a leave-one-out basis using 21 nearest neighbors and the Euclidean distance metric. In addition to the prediction, we can calculate an associated certainty measure, which represents the proportion of the nearest neighbors with the predominant mechanistic class. The 21 nearest neighbors can be ranked by similarity based on their Euclidean distance to the compound of interest.

Automated data processing

All routine data collation, normalization, and application of multivariate statistical techniques were incorporated into an automated data processing and reporting workflow created using the Pipeline Pilot software (Accelrys). This workflow was built using a set of core processing components designed specifically for handling multivariate plate-based high-content screening data and allows new workflows to be rapidly tailored to new assays. The workflow first joins raw data-relating compound and cell plates with results from the Definiens image analysis. Data from the different fields of view in each well are aggregated and then normalized to the plate controls. The Pipeline Pilot workflow is then used to automate the execution of the R algorithms to perform PCA, the Mahalanobis hit stratification, and kNN clustering.

Results

High-content image analysis of complex phenotypes across distinct cell types

The phenotypic response of cells following chemical perturbations is inherently unstable; multiple phenotypes may appear transiently with distinct temporal dynamics for any specific compound and dose. To define the optimal time point for our high-content phenotypic analysis, we used the brightfield kinetic imaging capabilities of the Cell-IQ instrument coupled with associated machine-learning algorithms to kinetically monitor defined phenotypic responses (Supplementary Fig. S1). Supplementary Fig. S1B and C shows how the kinetics

of the transient mitotic arrest phenotype induced with the microtubule-disrupting drug colchicine can vary dependent on cell type and dose. From the kinetic phenotypic response data obtained, 24- and 48-hour time points following compound exposure were selected for detailed high-content analysis to ensure the capture of a broad range of both transient and sustained phenotypic response across cell types.

The integrity of the actin and microtubule cytoskeleton is critical for cell viability and support cell motility, nuclear division, and cytokinesis that drives tumor growth and metastasis (21). We have developed a multiplex high-content cell morphology assay to monitor cytoskeletal, nuclear, and cellular morphology in detail across distinct, MCF7-wt (wild-type p53), MCF7-p53 (mutant p53), MiaPaCa2, and Ovcar3 cancer cell types.

To compare and contrast small molecule–induced change in morphology between relevant cancer cell types, our aim was to design a generic image analysis algorithm to automatically monitor both complex and subtle changes in cell morphology across distinct cancer cell types without any manual image-based thresholding. This challenge is further complicated by distinct heterogeneous cell morphologies in control and compound-treated cells across the four cell types selected for study (Fig. 1A and B). To manage the vast variety of phenotypes produced from screening mechanistically distinct compounds, a bespoke image analysis algorithm was designed using the Definiens Cellenger software package (22). Initially, the algorithm segments cells by accurately defining cellular and nuclear boundaries (see Materials and Methods). HCS CellMask whole-cell stain is used by the algorithm to define the boundary between a cell's cytoplasm, that of neighboring cells, and the background, thereby providing a robust cellular segmentation algorithm. When applying this method, even tightly packed cells within a population can be successfully segmented from one another (Fig. 1A). The algorithm uses the DAPI DNA-binding stain to define the nuclear border (Fig. 1A). Following cell and nuclear segmentation the algorithm next quantifies cytoskeletal and DNA marker intensities, shape, and texture as well as whole-cell shape metrics for each cell. As the Definiens algorithm is context based, every measurement can be related to each segmented cell. Due to the wide heterogeneity of cell phenotypes observed within cell populations especially following compound perturbation, a further level of subpopulation segmentation is beneficial (23, 24).

The subpopulation segmentation in this study was achieved through a rationale image analysis–based approach, using ratiometric analysis and automated thresholding of selected cell shape features and intensity measurements (see Materials and Methods). Using this method, cells were initially classified into round and other (nonelliptical) populations. A numerical threshold relating to the number and size of pseudopods was predefined to discriminate the pseudopod phenotype from a normal morphologic phenotype. Pseudopod phenotypic classifica-

tion corresponds to elongated cells that contain significantly increased numbers of filamentous actin containing pseudopod and lamellipod extensions compared with a cell classified as “normal” phenotype (see Materials and Methods for detailed subclassification criteria). Therefore, in this study, cells were segmented into four distinct subpopulations: round, other, pseudopod, and normal (Fig. 1C). Within each subpopulation, over 100 direct feature parameters can be extracted and quantified per cell (see Supplementary Fig. S3 for a complete list of parameters). Such classification enables detailed analysis of complex and heterogeneous phenotypes, and cellular subpopulations at the well level. Among multiple varying phenotypes, which can be detected using this algorithm, we also show the ability to profile epithelial to mesenchymal transitions (Fig. 1D). Incubation of MCF7-wt cells with a nonselective Src tyrosine kinase inhibitor, PP2, at 3 μmol/L for 24 hours induces an epithelial phenotype. This epithelial phenotype can be differentiated from the mesenchymal phenotype induced following incubation with the HDAC-1 and nitric oxide synthase inhibitor valproic acid at 150 μmol/L for 24 hours (Fig. 1D). The kinetic and multiparametric image analysis approaches described in Supplementary Fig. S1 and Fig. 1 enable a unified approach to the analysis of complex and distinct phenotypic responses across relevant cell models following compound treatment.

Automatic statistical analysis of multiparametric phenotypic data and clustering by MOA

The analysis of large volumes of multiparametric data across multidimensional assay formats incorporating time series and distinct cell types in a time frame suitable for drug discovery presents a significant logistical challenge. Underpinned by the Pipeline Pilot software from Accelrys, we have developed an innovative data analysis protocol incorporating multivariate statistical tools to automatically collate, analyze, and rank multiparametric high-content phenotypic results.

PCA enables simple visualization of large, complex, multiparametric phenotypic response across a dose response. Figure 2A shows the projection of the sample scores onto the first three principal components, which explain ~60% of the variation in the data set. The red cloud in the middle is the group of negative controls (0.1% DMSO-treated samples); each other point on the plots represents a compound at a particular dose. The points have been colored by compound and joined by lines between doses. Distinct phenotypic responses across a dose range are often displayed as distinct trajectories within the three-dimensional scattergram plot of three principal components (Fig. 2). The Mahalanobis distance metric is a multiparametric distance statistic (20), which can be used to define activity in an assay by identifying atypical responses either in shape or distance from the cloud of DMSO vehicle–negative controls (see Materials and Methods). A Mahalanobis distance threshold is calibrated for each assay by manually

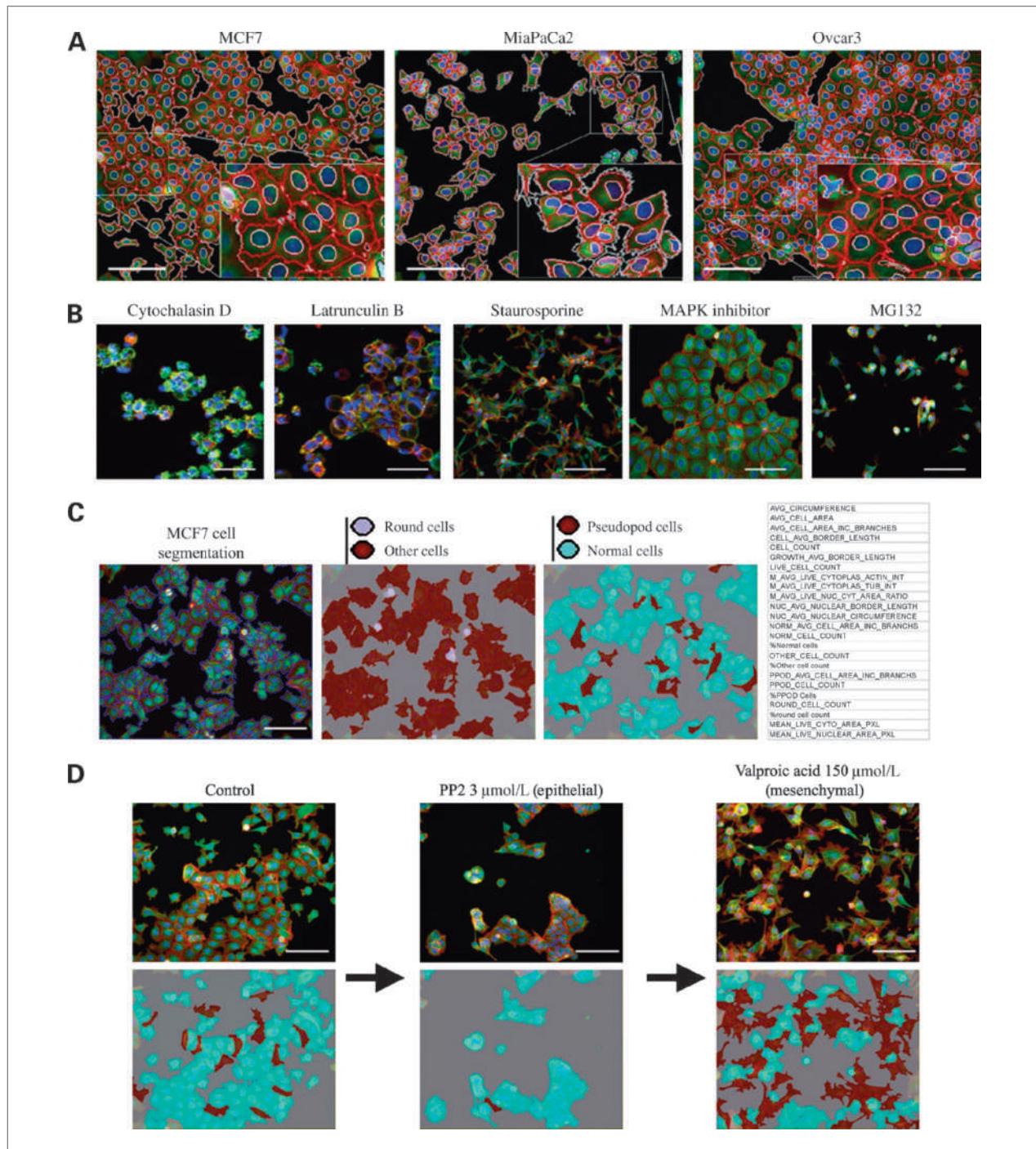


Figure 1. Bespoke image analysis across distinct cancer cell types. A cellular morphology assay was designed to monitor, in detail, changes in cytoskeletal and nuclear architecture across multiple cancer cell types. This four-wavelength multiplex assay consisted of the DNA binding dye, DAPI (blue), an actin cytoskeleton marker, Phalloidin (red), anti- β -tubulin antibody (green) to monitor the microtubule cytoskeleton, and the cytoplasmic marker, HCS cell mask. A, an image analysis algorithm created using Definiens Cellenger first segments individual cells by identifying nuclear and cytoplasmic boundaries. Cell and nuclear segmentation are indicated by red and white pseudocolor, respectively. B, examples of the broad spectrum of heterogenic phenotypes captured when applying this assay after compound incubation. Examples shown are from MCF7-wt cells. C, following cell segmentation, the Definiens image analysis algorithm captures distinct parameters for each individual cell, which represent a variety of geometric, intensity, subcellular localization and texture features. Based on a selection of feature parameters, the algorithm assesses changes across specific subpopulations, such as normal, round, and pseudopod (pseudopod class refers to elongated cells containing a higher proportion of filamentous actin-containing projections than normal cells). D, validation of this subpopulation analysis was done by monitoring transition to epithelial and mesenchymal phenotypes following treatment with PP2 and valproic acid, respectively. Scale bars, 100 μ m.

checking a selection of control and compound-treated images around the periphery of the DMSO control cloud, as indicated by PCA. The user-defined Mahalanobis distance metric and appropriate confidence limits ($P < 0.01$) are then automatically applied across all total compounds and doses to provide a robust and sensitive hit stratification method for multiparametric high-content data (Fig. 2A).

The PCA analysis presented in Fig. 2B shows the multiparametric analysis of our high-content morphology assay on compound-treated MCF7-wt cells. Protein synthesis inhibitors (anisomycin, cyclohexamide, emetine, and rapamycin) can be differentiated from the protease inhibitor class [ALLN (N-acetyl-Leucine-leucine-Norleucinyl), doxorubicin, proteasome inhibitor 1, MG132, and chloramphenicol]. This occurs despite the

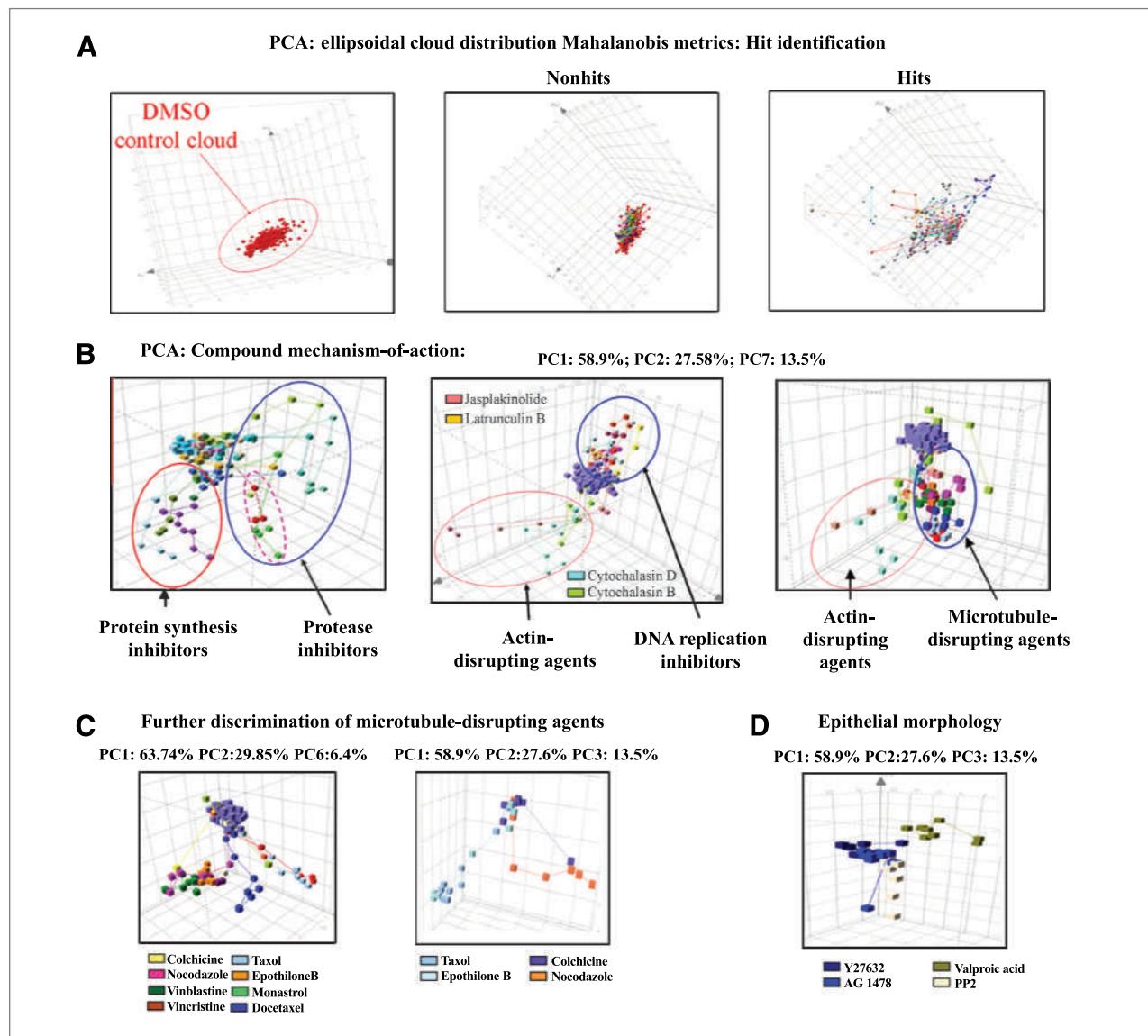


Figure 2. PCA incorporating Mahalanobis distances to identify active compounds. **A**, working in the principal components domain, the Mahalanobis distance metric is used to identify compound-induced phenotypes significantly different from the multiparametric DMSO control ellipsoid cloud distribution. Compounds producing atypical parameter responses from the DMSO cloud are identified and those above a calibrated Mahalanobis distance threshold are automatically stratified and classified as a "Hit" in the morphology assay. **B**, PCA visualization identifies distinct phenotypic responses across a dose response that relate to compound MOA such as inhibition of protein synthesis, disruption of actin filaments, DNA replication inhibitors, and microtubule-disrupting agents. In addition, target selectivity, structure activity relationships associated with particular phenotypic outcome, can be identified from within the protease inhibitor panel set (dotted lines, MG132 and ALLN). The Eigenvalues for all three principal components displayed are as follows: PC1, 58.9%; PC2, 27.58%; and PC7, 13.5%. **C**, further discrimination of MOA of microtubule-disrupting agents can be visualized by PCA. Eigenvalues for PCA plots are displayed. **D**, transition between epithelial and mesenchymal subpopulations can also be defined by PCA of the multiparametric image analysis measurements.

protease inhibitor class being composed of a diverse chemical set that portray a broad cluster of phenotypes over a large area of phenotypic space. Within this chemically diverse protease inhibitor set, evidence exists of target selectivity and structure activity relationships associated with distinct phenotypic outcome. The only two active peptide aldehyde inhibitors, MG132 and ALLN, cluster tightly and separately from the rest of this protease class (Fig. 2B). DNA replication inhibitors (aphidicolin, arabinofuranosylcytosine, filipin, floxuridine, hydroxyurea, and mitomycin C) cluster separately and can be differentiated from the set of actin-disrupting agents (cytochalasin B, cytochalasin D, latrunculin B, and jasplakinolide; Fig. 2B). Within the actin-disrupting agent cluster, jasplakinolide, while presenting a similar phenotypic tangent to latrunculin B and the cytochalasins B and D, also segments separately at higher doses, representing a distinct mode of action possibly due to a reported competition with a phalloidin binding site on filamentous actin (25). Eigenvalues for PCA scattergrams displayed in Fig. 2B are as follows: PC1, 58.9%; PC2, 27.58%; and PC7, 13.5%. PCA analysis not only distinguishes between actin and microtubule-disrupting agents but also can further discriminate between microtubule-disrupting and stabilizing agents based on heterogenous phenotypic response at the well level (Fig. 2C). Eigenvalues for PCA scattergrams displayed in Fig. 2C are as follows: PC1, 63.74%; PC2, 29.85%; PC6, 6.4% (microtubule disrupting agents); and PC1, 58.9%; PC2, 27.6%; and PC3, 13.5% (Taxol/epothiloneB versus colchicine/nocodazole). Epithelial to mesenchymal transition in cancer cells has significant relevance to tumor metastasis and therapeutic response. Using PCA analysis, we show that our Definiens algorithm and multiparametric high-content assay can differentiate epithelial (AG1478 and PP2) and mesenchymal (Y27632 and Valproic acid) phenotypes in MCF7-wt cells (Fig. 2D). Eigenvalues for PCA scattergrams displayed in Fig. 2C (epithelial versus mesenchymal) are as follows: PC1, 58.9%; PC2, 27.6%; and PC3, 13.5%. All PCA data presented in Fig. 2 are of MCF7-wt cells (data from other cell types are not shown). The automated Mahalanobis and PCA protocols run in pipeline pilot enable rapid and robust hit stratification and visualization of complex phenotypic responses across compound dose response. Reference to Eigenvalues and associated PCA loadings indicates precise phenotypic features that discriminate between compound-induced response, thereby informing compound MOA.

Kohonen networks: profiling phenotypic response across distinct cancer cell lines

To monitor distinct phenotypic responses across distinct cancer cell types, we have used a neural network approach (26). A Kohonen network (self-organizing map) was generated to display multidimensional phenotypic data for each compound across dose response and cell type in a two-dimensional grid format (26). The Kohonen network enables the direct visualization of distinct multiparametric phenotypic responses to a specific compound treatment across the panel of four cell lines

tested in this study (Fig. 3). When a Kohonen network has converged, data points that are similar to each other, in this case representing a similar multiparametric phenotypic response, will cluster together in similar areas of the map. By analyzing the topological relationships between points on the map, we can gain insight into the phenotypic relationships present in the data set. Within the reference library of compounds tested in this study, select compounds elicited a variety of resistance and sensitivity patterns across our cell panel, as determined by subsequent cross-reference of phenotypic space on the map to specific phenotypic features including cell number values and original images. In addition, some compounds produced the same phenotypic response across the four cell lines in which the data points over the dose response cluster together within the two-dimensional grid. Compounds, such as demecoline, show clustering across the dose response in accordance to cell line tissue origin (Fig. 3C); for example, MCF7-wt and p53 mutant cells cluster together, representing a characteristic microtubule disruption phenotype exemplified by nuclear fragmentation and diffuse microtubule staining (see Fig. 4A). MiaPaCa2 and Ovar3 cells cluster separately, representing a highly sensitive response as shown by reduced cell number and cytotoxic phenotype, and somewhat resistant phenotypes, respectively (Fig. 3C). Epothilone B, the microtubule stabilizer, induces a similar phenotypic response across all four cell lines tested (Fig. 3B). Following emetine treatment, the Ovar3 cells cluster separately from MCF7 and MiaPaCa 2 cells; furthermore, the phenotypic response of MCF7-wt cells is different from MCF7-p53 cells (Fig. 3D). Corresponding emetine phenotypic response pattern with cell number values and images indicate that MiaPaCa2 and MCF7-p53 cells are more sensitive to emetine treatment (as shown by reduced cell number and a cytotoxic appearance) relative to MCF7-wt and Ovar3 cells.

Kohonen networks can enable the rapid visualization of compound-induced phenotypic responses across multiple dimensions including dose-response and cell type. Corresponding patterns of phenotypic response across cell types to cytotoxic phenotypic characteristics and cell type-specific mutation status can rapidly elucidate therapeutic sensitivity and resistance across cell models, as well as uncover novel mechanistic insight in relation to genetic background and mutation status (e.g., p53 mutation status as shown in Fig. 3D).

kNN statistics: clustering and ranking compounds by biological MOA and uncovering novel mechanistic insight

The kNN algorithm is a simple but intuitive algorithm that classifies new objects from unknown groups based on their phenotypic similarity to that of an annotated training set. It starts with multivariate measurements on the training set, for which the correct classifications (MOA) are already known. For an object with an unknown classification, we first determine which of the

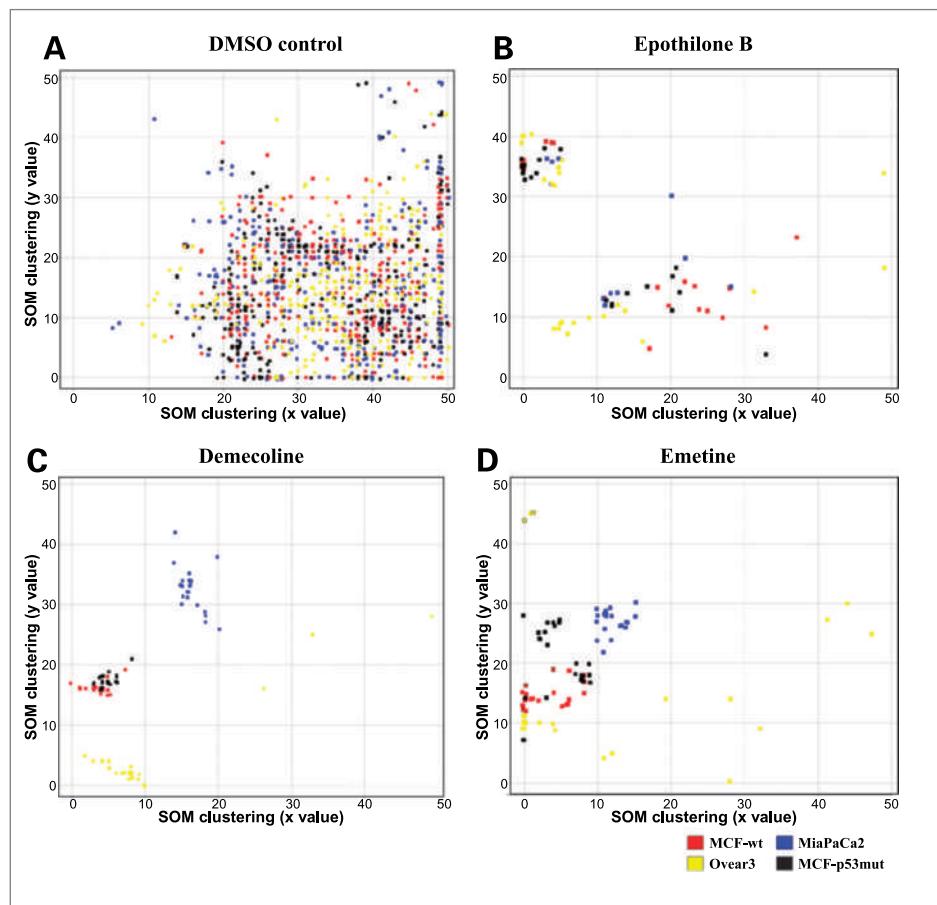


Figure 3. Kohonen networks: profiling phenotypic response across distinct cancer cell lines. A Kohonen network was generated to display multiparametric image analysis data representing distinct dose-dependent phenotypic responses across multiple cancer cell types. Each data point represents the multiparametric phenotype measured by the Definiens image analysis algorithm for each specific dose of tested compounds. Data shown displays phenotypic distribution across DMSO controls (A), dose response of epothilone B (B), demecoline (C), and emetine-treated MCF7-wt, MCF7-p53-mut, MiaPaCa2, and Ovar3 cells (D).

training set are closest (most similar) to the new object and then assign the new object to the class that is most common among its k -nearest neighbors. kNN allows quantitative clustering and the MOA of test compounds to be predicted based on their phenotypic similarity to those of an annotated reference set of compounds as defined by the multiparametric Definiens analysis of the morphology assay. Figure 4 shows a sample output from an interactive kNN classifier tool, which was integrated with the high-content analysis data and the Mahalanobis hit-stratification method using Accelrys' Pipeline Pilot software. This tool enables users to select any active compound or dose from the high-content assay and then run the kNN classifier across the entire reference set. The output displays the selected compound of interest, its dose, and predicted MOA based on the most prevalent MOA among its 21 nearest neighbors from the reference set. A probability score for the assigned MOA is calculated by the proportion of its 21 nearest neighbors representing the predominant MOA.

In addition, the 21 nearest compounds or neighbors are ranked in order according to their similarity as determined by Euclidean distance metrics, the shorter the distance the more similar the phenotype (Fig. 4A and B).

When cross-referencing the phenotypic signature of MCF7-wt cells after 24 hours of exposure to 1 $\mu\text{mol/L}$ demecoline, a microtubule-disrupting agent, the kNN tool predicts with a 95% KNNCV probability that demecoline is in fact a microtubule-acting compound (Fig. 4A). The 13 nearest neighbors to demecoline are specifically microtubule-disrupting agents (Fig. 4A). Raw acquisition images of control MCF7-wt cells and 24 hours following exposure to 1 $\mu\text{mol/L}$ demecoline and 0.3 $\mu\text{mol/L}$ vincristine (nearest neighbor to 1 $\mu\text{mol/L}$ demecoline) are shown to validate the interactive kNN results (Fig. 4A). Similarly, for the microtubule stabilizer paclitaxel's phenotypic readout at 0.03 $\mu\text{mol/L}$, the seven nearest neighbors are also microtubule-stabilizing agents (Fig. 4B).

The Euclidean distance of nearest neighbors can be used to further compare compound MOA across the panel of cell lines. Table 1 displays the top five nearest neighbors, their dose (enclosed brackets), and Euclidean distance similarity metric (enclosed asterisk) for seven selected compounds at a chosen active dose across the four cell types based on the Definiens high-content morphology study (24-h time point following compound treatment).

For demecoline at 1 $\mu\text{mol/L}$, the closest neighbor is vincristine followed by a range of vincristine doses with

short Euclidean distances (high degree of similarity) across all cell types. This suggests that demecoline and vincristine share a similar MOA (microtubule destabilizers) within each cell line. However, with the aid of Kohonen networks, we deduce that demecoline has a different phenotypic response and potential cytotoxic potency between the tissue origins of the cell panel (Supplementary Fig. S5; Fig. 3). Phenotypic analysis of camptothecin (a

topoisomerase I inhibitor and DNA damage–causing agent) at 1.5 μmol/L shows that its nearest neighbors are exclusively DNA-damaging or DNA replication-inhibiting agents across all cell types. These results show the capability of our assay and associated kNN tool to identify DNA damage and DNA-repairing agents. Interestingly, the five nearest neighbors for MCF7-wt and MCF7-p53 cells are flouxuridine, mitomycin C, and

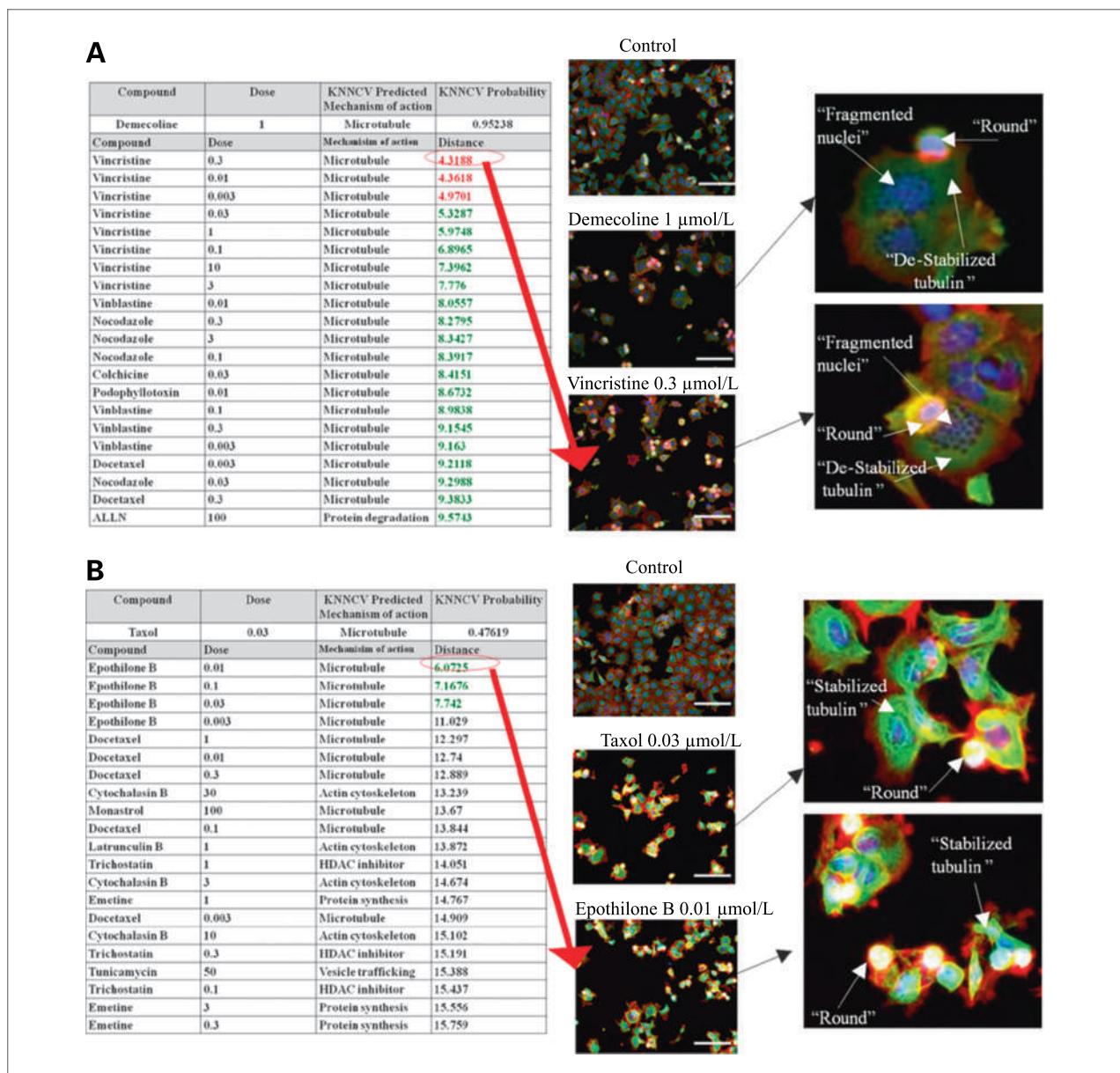


Figure 4. kNN clustering by MOA. A kNN nearest neighbor classifier was incorporated into an automatic interactive user interface to identify similar phenotypic response and predict MOA based on similarity between multiparametric image signatures of test compounds with that of a reference library of compounds. A, output of kNN analysis (MCF7-wt cells) to 1 μmol/L Demecoline (24 h). B, output of kNN analysis to 0.03 μmol/L paclitaxel (24 h). The user interface displays the tested compound and its active dose; it also displays the compound's predicted MOA and the probability of this based on the prevalent MOA of its 21 nearest neighbors. It next displays and ranks similarity based on the Euclidean distance from the selected test compounds: the smaller the distance, the closer in phenotypic readout to the test compound. Associated images were collected to validate results. Scale bars, 100 μm.

Table 1. kNN clustering: Profiling mechanism of action across cell types

	1	2	3	4	5
Demecoline 1μM					
MCF7-wt	Vincristine (0.3) *4.31*	Vincristine (0.01) *4.36*	Vincristine (0.003) *4.97*	Vincristine (0.03) *5.32*	Vincristine (1) *5.97*
MCF7-p53	Vincristine (0.03) *6.02*	Vincristine (0.1) *7.25*	Vincristine (0.01) *7.27*	Vincristine (0.3) *7.72*	Vincristine (3) *7.72*
MiaPaCa2	Vincristine (0.3) *6.89*	Vincristine (0.1) *7.28*	Vincristine (0.01) *7.48*	Vincristine (0.03) *7.54*	Vincristine (3) *8.30*
OvCar3	Vincristine (0.3) *4.14*	Vincristine (10) *4.21*	Vincristine (1) *4.37*	Vincristine (0.01) *5.23*	Vincristine (0.03) *5.71*
Bleomycin 1.5μM					
MCF7-wt	Doxorubicin (0.03) *4.49*	Etoposide (1) *4.83*	Mitoxantrone (0.003) *4.97*	Forskolin (0.3) *4.98*	Etoposide (0.3) *5.31*
MCF7-p53	Lactacystatin (10) *4.98*	Doxorubicin (0.01) *5.02*	Alosine A (30) *5.05*	Camtotoxin (3) *5.66*	Quercetin (3) *5.66*
MiaPaCa2	Aphidicolin (1) *6.19*	Chlorambucil (10) *6.38*	Aphidicolin (10) *6.58*	Aphidicolin (0.1) *6.60*	Aphidicolin (3) *7.01*
OvCar3	Sodium Fluoride (15) *8.07*	Sodium Fluoride (0.15) *9.10*	Trichostatin (0.03) *9.32*	Sodium Fluoride (50) *9.49*	Hydroxyurea (10) *9.65*
Camptothecin 0.0015μM					
MCF7-wt	Flouxuridine (3) *4.45*	Flouxuridine (30) *5.30*	Mitoxantrone (0.01) *5.39*	Mitomycin C (3) *5.39*	Mitomycin C (0.1) *5.41*
MCF7-p53	Flouxuridine (3) *6.14*	Flouxuridine (100) *6.15*	Mitomycin C (1) *6.32*	Flouxuridine (30) *6.42*	Mitoxantromine (0.01) *6.55*
MiaPaCa2	Etoposide (3) *4.15*	Aphidicolin (10) *5.25*	Aphidicolin (3) *5.49*	Etoposide (1) *5.53*	Etoposide (30) *5.54*
OvCar3	Aphidicolin (0.1) *5.45*	Etoposide (30) *6.13*	Aphidicolin (1) *6.21*	Etoposide (10) *6.24*	Aphidicolin (0.3) *6.28*
Proteasome Inhibitor 1 10μM					
MCF7-wt	H-7 (50) *9.66*	Okadaic Acid (0.02) *9.81*	MG132 (10) *9.93*	ALLN (10) *10.12*	MG132 (1) *10.16*
MCF7-p53	MG132 (1) *10.35*	MG132 (3) *10.82*	H-7 (5) *11.58*	ALLN (10) *12.31*	ALLN (100) *9.73*
MiaPaCa2	MG132 (1) *8.23*	MG132 (3) *8.73*	ALLN (10) *8.94*	ALLN (10) *9.55*	ALLN (100) *12.74*
OvCar3	PP2 (10) *8.41*	Anisomycin (0.3) *8.71*	Anisomycin (3) *9.43*	cycloheximide (5) *9.50*	PP2 (3) *9.65*
Emetine 1μM					
MCF7-wt	cycloheximide (50) *5.96*	Tunicamycin (50) *6.04*	Anisomycin (3) *7.09*	Cycloheximide (15) *7.22*	Anisomycin (0.3) *7.50*
MCF7-p53	Anisomycin (10) *6.04*	Anisomycin (3) *6.70*	Cytochalasin D (0.3) *7.70*	Cytochalasin B (3) *8.20*	Anisomycin (1) *8.40*
MiaPaCa2	cycloheximide (15) *6.93*	Tunicamycin (15) *7.23*	Tunicamycin (5) *7.29*	Tunicamycin (0.5) *7.34*	cycloheximide (5) *7.62*
OvCar3	Brefeldin A (0.3) *7.36*	Okadaic Acid (0.02) *8.19*	Cisplatin (30) *8.31	Brefeldin A (0.1) *8.55*	Brefeldin A (1) *8.70*
Epothilone B 0.1μM					
MCF7-wt	Paclitaxel (0.03) *7.10*	Paclitaxel (0.3) *7.50*	Paclitaxel (1) *7.72*	Paclitaxel (3) *8.87*	Paclitaxel (0.1) *8.95*
MCF7-p53	Paclitaxel (0.3) *7.01*	Paclitaxel (0.01) *7.11*	Paclitaxel (0.1) *7.33*	Paclitaxel (0.03) *7.62*	Paclitaxel (3) *7.71*
MiaPaCa2	Ocytochalasin B (3) *31.58*	Staurosporine (0.3) *32.28*	Paclitaxel (0.0) *33.44*	Paclitaxel (0.001) *34.04*	Paclitaxel (0.03) *34.19*
OvCar3	Paclitaxel (0.03) *6.70*	Paclitaxel (0.1) *6.72*	Monastrol (100) *7.30*	Paclitaxel (0.3) *7.36*	Paclitaxel (0.01) *7.55*
Lovastatin 5μM					
MCF7-wt	Simvastatin (20) *6.02*	Simvastatin (6) *7.78*	H-7 (15) *10.53*	Trichostatin (0.03) *10.80*	MG132 (30) *10.96*
MCF7-p53	Simvastatin (6) *8.11*	H-7 (5) *10.00*	Trichostatin (0.03) *10.04*	U0126 (30) *10.06*	Alosine A (30) *10.57*
MiaPaCa2	Simvastatin (6) *6.32*	Brefeldin A (0.03) *7.07*	Simvastatin (20) *8.83*	Floxuridine (1) *8.86*	Floxuridine (0.3) *8.88*
OvCar3	PP2 (0.1) *4.29*	Diaminobenzidine (10) *4.34*	Simvastatin (2) *4.87*	cycloheximide (0.5) *5.40*	Olomoucine (10) *5.43*
Colour Key MOA	Actin cytoskeleton	Kinase (broad)	protein degradation	Vesicle trafficking/Glycosylation inhibitors	
Microtubule modifier	DNA replication/damage	HDAC inhibitors	Kinase (CDK)	Statin-lipid	
			Kinase synthesis	Kinase (MAPK)	

Concentration: (μM within parenthesis); ** denotes Euclidean distance.

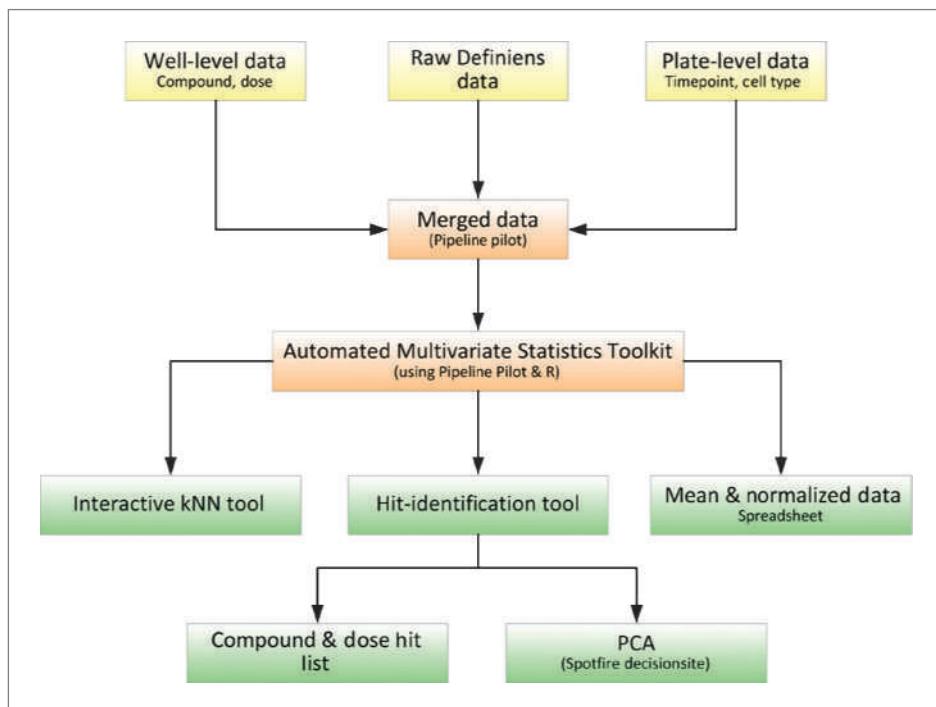


Figure 5. Automated high-content phenotypic data processing workflow. Using the Pipeline Pilot software, a series of data processing scripts are linked to automate the aggregation, normalization, and collation of Definiens raw image analysis data with meta-data describing experimental parameters. The pipeline then integrates the merged data with the application of multivariate statistical analysis algorithms created using the R software including Mahalanobis hit stratification, PCA, and kNN classification to produce several reports and interactive html files that enable rapid interrogation of multiparametric phenotypic response data.

mitoxantrone, whereas those for MiaPaCa2 and Ovar3 cells are etoposide and aphidicolin (Table 1). These results indicate that camptothecin and/or its nearest neighbors display modified MOA across different cell types. Another DNA-damaging agent, bleomycin, at 1.5 μmol/L exhibits a much more diverse set of nearest neighbors across the cell panel. Here, all of the top five nearest neighbors of MiaPaCa2 cell are DNA replication or DNA-damaging agents and the nearest neighbors of MCF7wt cell line are also predominantly within this class of compounds. In contrast, the predominant nearest neighbor of Ovar3 cell is sodium fluoride and the closest compounds of MCF7-p53 mutant cell are made up of a range of classes with no one particular mechanism being prevalent (Table 1). This variety in mechanistic similarity may be because in contrast to specific targeted inhibitors, Bleomycin has an effect upon multiple mechanisms that can be influenced by genetic and epigenetic factors specific to certain cell backgrounds.

Following treatment with the protein synthesis inhibitor Emetine at 1 μmol/L, MCF7-wt, MCF7-p53 mutant, and MiaPaCa2 cells indeed have protein synthesis inhibitors as their closest neighbors (Table 1). In contrast, for Ovar3 cells, there is no appearance of protein synthesis inhibitors as emetine's nearest neighbors, suggesting that incubation with emetine for 24 hours has no specific effect on protein synthesis at 1 μmol/L or has a different MOA to other protein synthesis compounds within our training set in this cell line. The actin disrupters cytochalasin D and B appear as nearest neighbors to emetine at 1 μmol/L, in MCF7-p53 cells

(Table 1). Indeed, increased evidence of actin disruption and a higher proportion of round cells can be observed in raw images (Supplementary Fig. S6). This provides some evidence that emetine may have a distinct influence upon cytoskeletal and nuclear morphology dependent on the p53 status within MCF7 cells. This hypothesis is further strengthened when analyzing emetine with Kohonen networks as MCF7-wt and MCF7-p53 mutant cells cluster separately from each other as shown in Fig. 3.

Although simvastatin and lovastatin are the only two lipid-lowering compounds within our reference library, our cytoskeletal morphology assay and methods are powerful enough to cluster them closely based on the similarity of their phenotypic fingerprint. When profiling lovastatin at 5 μmol/L, simvastatin is its closest neighbor in all cell lines except Ovar3 in which it is third closest; however, here, simvastatin at 2 μmol/L has the smallest Euclidean distance (4.87) across cell types, indicating a similar phenotype to lovastatin.

We show how the application of the kNN tool can predict the mechanism-of-action of a test compound by reference to an annotated training set. We further show how our kNN interactive tool incorporating Euclidean distance metrics can rank compound similarity to a compound of interest and thus facilitate compound selection and decision making based on desired phenotypic response. When applied to our multiparametric high-content morphology assay data, we show the ability to monitor differences in compound MOA across cell types. We further show the sensitivity and robustness of the

approach by identifying MOA based on both subtle and complex heterogenous phenotypic changes analyzed at the well level.

Interactive automation capabilities for high-throughput data analysis

A schematic of the fully automated data analysis workflow created and implemented during this piece of work using Accelrys's Pipeline Pilot software is shown in Fig. 5. The raw data acquired from the Definiens algorithm is initially collated with the relevant compound, dose, cell type, and time point. Data from the four fields of view in each well is then aggregated before being normalized to the median of the negative DMSO controls. Once data are normalized, the pipeline creates multiple output files including a Microsoft Excel file of the raw data, PCA visualization in Spotfire, a Mahalanobis hit stratification tool resulting in an active dose "hit" identification list, and the interactive kNN html tool. The use of this automated Pipeline Pilot workflow enables the rapid and robust application of the multivariate analysis and statistical tools described above and, in doing so, increase the quality, consistency, and integrity of the data. This automated system empowers scientists to routinely use high-end statistical and computational analysis to visualize and quantify complex drug-induced phenotypes in a timeline suitable for high-throughput screening.

Discussion

A primary objective of image-based high content technologies are to provide a more informative and physiologically relevant screening platform for selecting high-quality drug candidates for further development (3, 8, 27). It is anticipated that such methods shall reduce high levels of compound attrition, during later stages of discovery and development, by prioritizing new classes of therapeutics with improved efficacy and toxicity profiles (27). To achieve such predictive power, it is necessary that high-content assays recapitulate the complexities of *in vivo* biology, including distinct genetic and epigenetic variants that represent the broad spectrum of physiology and disease pathology exhibited across heterogeneous patient populations and evolving disease pathophysiology. In this article, we describe a flexible custom-designed high-content phenotypic approach integrated into an automated data processing workflow to rapidly and robustly monitor phenotypic response across physiologically relevant cell panels that either reflect patient mutation status or distinct cancer origin. We describe the application of a series of novel image analysis and data processing tools integrated into a standard high-content phenotypic profiling workflow to facilitate compound prioritization for further development based on relevant phenotypic readouts. The high-content workflow described in this article begins with the application of automated time-lapse phase-contrast microscopy and machine learning algorithms to define appropriate time points for the detailed moni-

toring of cellular phenotypic response. The creation of custom-designed image analysis algorithms with Definiens cognition network technology that yield the measurement of features across cellular subpopulations enables the capture of both complex and subtle phenotypic response across relevant cell types. The Mahalanobis metric provides a robust hit stratification method based on multiparametric phenotypic responses. Manual calibration of a Mahalanobis distance threshold to define active response from DMSO control followed by automated application to an entire data set enables robust identification of compound activity across broad phenotypic space for any compound, dose, and time point tested. Such a statistically driven approach to hit identification offers substantial advantages over the laborious visual examination of images, not only in terms of objectivity, reproducibility, and time saved by scientists but also in terms of sensitivity. From our experience, the Mahalanobis hit stratification tool is able to detect compound-induced phenotypes that represent subtle changes in cell morphology that may have been missed by the human eye. Generation of Kohonen networks enable simple visualization of complex multidimensional data sets. In this study, we apply a Kohonen network strategy to evaluate multiparametric phenotypic response signatures across dose concentrations and cell types. The application of Kohonen network to compare phenotypic response across distinct cell types is only achievable as a consequence of our development of a generic custom-designed image analysis algorithm that can, automatically, measure high-quality morphologic parameters across distinct cell types. kNN (nearest neighbor) classifier aligned to a Euclidean distance metric was applied in this study to rank compounds based on phenotypic similarity with an annotated reference set and then ascribe a predicted MOA and associated probability score for any selected compound.

In this study, we show how the application of Kohonen and kNN methods to multiparametric high-content data can uncover novel mechanistic insights by correlating compound-induced phenotypic response with cell type and genetic background. For example, using Kohonen network analysis, we show that emetine induced a distinct phenotypic response in MCF7-wt cells compared with MCF7-p53 mutant cells. These results suggest that emetine-induced effects upon the actin cytoskeleton of cancer cell lines may be dependent on p53 function. To our knowledge this is the first report describing such a p53 requirement for emetine-induced cellular function. Indeed, previous studies have concluded that p53 does not play a role in the emetine-induced apoptosis of tumor cell lines (28). The identification of novel p53 requirements for compound-induced changes in cytoskeletal architecture may have important consequences for tumor invasion and cytokinesis. In addition, Kohonen and kNN analysis highlight several distinct phenotypic responses between Ovcar3 cells and the other cell types. Manual visualization of images suggests that Ovcar3 cells are more resistant to the effects of several mechanistic

compound classes when compared with other cell types. Comparatively, MiaPaCa2 cells seem more sensitive to compound perturbation in many cases. The results from the kNN analysis presented in this study further validate the utility of the methods and workflow used by showing the ability to automatically define biological similarity within distinct mechanistic classes of compounds including microtubule-disrupting and microtubule-stabilizing agents, protease inhibitors, DNA replication inhibitors, and statins.

The application of high-content phenotypic profiling to relevant cell types and phenotypes described in this study enables clear clinical hypotheses to be tested, ensuring the collection of high-content results with biological meaning, thereby facilitating compound selection. We further show how these methods can be integrated into a generic data processing workflow to allow rapid decision making based on chemical and biological similarity.

As shown in this article, the ability to custom design Definiens image analysis algorithms to capture user-defined phenotypes across challenging cell types and heterogeneous responses indicates the potential to apply these approaches to primary, patient-derived tumor cells. Such methods may provide direct correlation between

phenotypic outcomes following compound treatment with heterogeneous tumor cell types including cancer-derived stem cell populations.

We anticipate that the application of the approaches described in this article aligned with physiologically relevant high-content screens offers the potential to drive the development of a new generation of therapeutics encompassing pharmacologic compositions or polypharmacology tailored to appropriate disease context and cellular subtype.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Andy Hargreaves, Emma Cooke, Lisa Drew, Christopher Denz, Lucienne Ronco, Lihua Yu, and Zhongwu Lai for the helpful discussions and advise during the course of these studies.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 12/10/2009; revised 03/10/2010; accepted 04/05/2010; published OnlineFirst 06/08/2010.

References

- Sams-Dodd F. Target-based drug discovery: Is something wrong? *Drug Discov Today* 2005;10:139–47.
- Butcher EC. Can cell systems biology rescue drug discovery? *Nat Rev Drug Discov* 2005;4:461–7.
- Bullen A. Microscopic imaging techniques for drug discovery. *Nat Rev Drug Discov* 2008;7:54–67.
- Lang P, Yeow K, Nichols A, Scheer A. Cellular imaging in drug discovery. *Nat Rev Drug Discov* 2006;5:343–56.
- Yarrow JC, Lechner T, Li R, Mitchison TJ. Rapid de-localization of actin leading edge components with BDM treatment. *BMC Cell Biol* 2003;3:4–5.
- Low J, Stancato L, Lee J, Sutherland JJ. Prioritizing hits from phenotypic high-content screens. *Curr Opin Drug Discov Dev* 2008;11:338–45.
- Tanaka M, Bateman R, Rauh D, et al. An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol* 2005;3:e128.
- Perlmutter ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. *Science* 2004;306:1194–8.
- Loo LH, Wu LF, Altschuler SJ. Image-based multivariate profiling of drug response from single cells. *Nat Methods* 2007;6:759–65.
- Young DW, Bender A, Hoyt J, et al. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat Chem Biol* 2008;4:59–68.
- Zhou BB, Zhang H, Damelin M, Geles KG, Grindley JC, Dirks PB. Tumour-initiating cells: challenges and opportunities for anticancer drug discovery. *Nat Rev Drug Discov* 2009;8:806–23.
- Ashworth A. Drug resistance caused by reversion mutation. *Cancer Res* 2008;68:10021–3.
- Bergers G, Hanahan D. Modes of resistance to anti-angiogenic therapy. *Nat Rev Cancer* 2008;8:592–603.
- Potti A, Dressman HK, Bild A, et al. Genomic signatures to guide the use of chemotherapeutics. *Nat Med* 2006;12:1294–300.
- Blick T, Widodo E, Hugo H, et al. Epithelial mesenchymal transition traits in human breast cancer cell lines. *Clin Exp Metastasis* 2008;25:629–42.
- Sahai E, Marshall CJ. Differing modes of tumour cell invasion have distinct requirements for Rho/ROCK signalling and extracellular proteolysis. *Nat Cell Biol* 2003;5:711–9.
- Denner P, Schmalowsky J, Prechtel S. High-content analysis in preclinical drug discovery. *Comb Chem High Throughput Screen* 2008;11:216–30.
- Gasperri F, Cappella P, Galvani A. Multiparametric cell cycle analysis by automated microscopy. *J Biomol Screen* 2006;11:586–98.
- Zhou X, Wong STC. Informatics challenges of high-throughput microscopy. *IEEE Signal Process Mag* 2006;23:63–4.
- Mahalanobis PC. On the generalised distance in statistics. *Proc Natl Inst Sci India* 1936;2:49–55.
- Hall A. The cytoskeleton and cancer. *Cancer Met Rev* 2009;28:1–2, 5–14.
- Baatz M, Arini N, Schape A, Binnig G, Linssen B. Object-oriented image analysis for high content screening: detailed quantification of cells and sub cellular structures with the cellenger software. *Cytometry Part A* 2006;69:652–8.
- Loo LH, Lin HJ, Steininger RJ III, Wang Y, Wu LF, Altschuler SJ. An approach for extensibly profiling the molecular states of cellular subpopulations. *Nat Methods* 2009;6:759–65.
- Low J, Huang S, Blosser W, et al. High-content imaging characterization of cell cycle therapeutics through *in vitro* and *in vivo* subpopulation analysis. *Mol Cancer Ther* 2008;7:2455–63.
- Bubb MR, Senderowicz AM, Sausville EA, et al. Jasplakinolide, a cytotoxic natural product, induces actin polymerization and competitively inhibits the binding of phalloidin to F-actin. *J Biol Chem* 1994;269:14869–71.
- Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybernet* 1982;43:59–69.
- Feng Y, Mitchison TJ, Bender A, Young DW, Tallarico JA. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat Rev Drug Discov* 2009;8:567–78.
- Moller M, Wink M. Characteristics of apoptosis induction by the alkaloid emetine in human tumour cell lines. *Planta Med* 2007;73:1389–96.