An integrated one-step system to extract, analyze and annotate all relevant information from image-based cell screening of chemical libraries†

Obdulia Rabal, Wolfgang Link, Beatriz G. Serelde, James R. Bischoff and Julen Oyarzabal*

Received 24th September 2009, Accepted 24th November 2009
First published as an Advance Article on the web 21st January 2010
DOI: 10.1039/b919830j

Here we report the development and validation of a complete solution to manage and analyze the data produced by image-based phenotypic screening campaigns of small-molecule libraries. In one step initial crude images are analyzed for multiple cytological features, statistical analysis is performed and molecules that produce the desired phenotypic profile are identified. A naïve Bayes classifier, integrating chemical and phenotypic spaces, is built and utilized during the process to assess those images initially classified as "fuzzy"—an automated iterative feedback tuning. Simultaneously, all this information is directly annotated in a relational database containing the chemical data. This novel fully automated method was validated by conducting a re-analysis of results from a high-content screening campaign involving 33 992 molecules used to identify inhibitors of the PI3K/Akt signaling pathway. Ninety-two percent of confirmed hits identified by the conventional multistep analysis method were identified using this integrated *one-step* system as well as 40 new hits, 14.9% of the total, originally false negatives. Ninety-six percent of true negatives were properly recognized too. A web-based access to the database, with customizable data retrieval and visualization tools, facilitates the posterior analysis of annotated cytological features which allows identification of additional phenotypic profiles; thus, further analysis of original crude images is not required.

Introduction

High-content screening (HCS) refers to an automated image-based analysis of intracellular events and cellular morphology by the treatment of cells with small molecules (or other agents).[1] This technology is increasingly being used for both primary and secondary screening assays in drug discovery.[2–6] Thus, multiple features of compound-induced cellular phenotypes that can be relevant from either a therapeutic or toxicological point of view are captured for many more compounds than in the past thereby enhancing the ability to select high-quality hits with the desired mechanism of action earlier in the discovery phase of a given project.

Image analysis and image processing techniques are essential to extract information from large scale cell-based assays. Commercial HCS automated microscopes have integrated the image acquisition and image analysis software,[7,8] although often these systems are not flexible enough for user customization. The posterior analysis of image-based phenotypic data sets requires time intensive user intervention such as the development and execution of specific scripts to convert and parse data files as well as performing statistical analysis. As a result, there is considerable interest for automated and adaptable solutions to mine biological data that are generated by high-throughput phenotypic screens that include integrated platforms of software for statistical analysis, methods to train a computer to score unusual cell morphologies automatically and network access to information and databases.[9–11]

In order to exploit the wealth of phenotypic information collected in the HCS data set the data must be captured and deposited in a database (DB) where chemical and biological knowledge are integrated. This requires efficient tools not only to integrate knowledge from these two data sets, but also to manage, handle and retrieve the information since this information is key for the iterative nature of the decision-making process required for drug discovery. The usual procedure to process data generated in HCS involves a time consuming sequential process of independent tasks including image analysis, retrieval of cytological features, statistical analysis, hit assessment, data formatting and finally annotation in a database with the corresponding chemical data. A further potential complication can be subjective user criteria which could be applied at each step.

We report here the development and validation of an integrated *one-step* method, completely user independent, that streamlines the data flow from captured images to an annotated database where cytological features, phenotypic profiles as well as hit assessment are integrated together with the corresponding chemical information. This DB is available

Experimental Therapeutics Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernandez Almagro, 3, 28029 Madrid, Spain. E-mail: joyarzabal@cnio.es; Tel: +34 91 7328000
† Electronic supplementary information (ESI) available: Details on collected cellular features, results from classical approach and statistical analyses (Student's *t*-test and confusion matrix). See DOI: 10.1039/b919830j

for the drug discovery team, medicinal chemists and biologists, through a flexible user-friendly web-based interface. To our knowledge, this is the first description of a *one-step* fully automated and systematic high-content analysis (HCA) method for image-based cell screening of small-molecule libraries.

Results

High-content screening assay

The HCS assay reported in this manuscript was designed to identify inhibitors of the PI3K/Akt signaling pathway. Type I phosphoinositide 3-kinases, such as PI3Kα, are involved in several fundamental biological pathways including cell survival, growth and differentiation as well as immunological responses. The importance of this pathway in human cancer is underscored by the fact that this pathway is activated in more than 50% of human cancers.[12]

The assay follows the Akt-dependent nuclear translocation of the FOXO family of transcription factors. Akt phosphorylates FOXO protein and induces a nuclear to cytoplasmic translocation. In a previous report,[13] we described a 96-well high-content imaging assay using U2OS osteosarcoma cells that stably express a GFP-tagged FOXO3a protein (U2Fox-Reloc). The GFP–FOXO translocation response was quantified as the relative distribution of the fluorescent GFP probe (Ch2) between the cytoplasm and the nucleus. Standard DNA staining (DAPI) was used as the fluorescent probe (Ch1) for individual cell segmentation, as a nuclear mask. We used this HCS to screen a collection of 33 992 compounds[14] (more details about HCS assay can be found in the Experimental section).

Image analysis and hit assessment: classical approach *via* BD Pathway Bioimager

Initially, for both assay development and compound screening, we used the image analysis routines available within the BD Pathway Bioimager through the AttoVision software and the Image Data Explorer for data processing (detailed information is provided in the Experimental section). From both, the DAPI and GFP channels a total of 11 cytological features (a table containing this information is available as ESI†, Table S1) were collected and exported as raw text files that were processed by Image Data Explorer, a custom Microsoft Excel based data analysis tool. We focused on cytological features that provide key information to assess the phenotype of interest, *i.e.* nuclear (FOXO) translocation,[13,14] and therefore facilitate the identification of compounds that induce the desired phenotypic profile (Hit Identification). For standard cell-based translocation assays, these features correspond to those that enable the quantification of the fluorescent probe between the cytoplasm and the nucleus. Visual inspection of images generated by the BD Pathway Bioimager was used to qualitatively confirm quantitative data from the data analysis process.

The data were uploaded into a DB which linked the chemical information with the retrieved cytological features as well as with the corresponding phenotypic profiling and the hit assessment.

Image analysis and hit assessment: novel *one-shot* approach

In order to automate the image analysis and hit assessment of an HCS campaign in a systematic manner a new tool was developed. The tool was implemented within a Pipeline Pilot platform,[15] in order to take advantage of the platform's customizable components for image analysis, data and workflow integration, reporting and so on. Once this was completed, a validation process was performed using the image files (TIF format) from a screen of 33 992 compounds.[14] The images, exported by the BD Pathway Bioimager, were re-processed and the data were re-assessed using the automatized *one-step* method.

Image analysis

First, a cytoplasm-to-nuclear translocation algorithm was implemented to quantify the relative distribution of GFP between the cytoplasm and the nucleus. The first step of the cytoplasm-to-nuclear translocation algorithm uses the White Top Hat[16] by Reconstruction and the Otsu segmentation[17] method to optimally threshold the image acquired in the DAPI channel. Objects touching the image boundary are removed. Next, nuclear peaks are found and used as markers for watershed segmentation to identify and label the nuclear regions of interest (nuclear masks). Nuclear masks having an

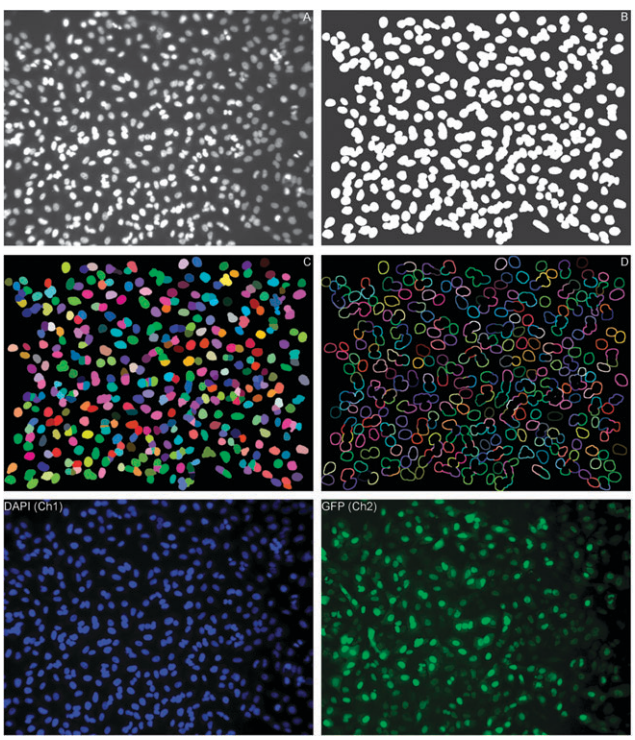

Fig. 1 Different stages of the image analysis carried out by this new tool. (A) Original 12-bit DAPI image, (B) binary image resulting of nuclear segmentation, (C) nuclear objects as labeled by Watershed from markers, a random pseudo color component was applied for clarity when depicting the original gray-scaled image. (D) Ring objects labeled to identify cytoplasmic regions of interest as obtained from a second watershed from markers using the dilated binary segmented nuclear mask. As for image (C), a random pseudo color component was applied for displaying purposes. Nuclear and cytoplasmic masks are applied to the original DAPI (Ch1) and GFP (Ch2) images to acquire intensity- and shape-related features.

area smaller than 100 pixels are removed, while the remaining are dilated 10 pixels to cover the cytoplasmic region. Watershed from markers component, using the aforementioned nuclear peaks, is applied to the resulting dilated image to label the nuclear plus cytoplasmic regions (Fig. 1). The nuclear and the nuclear plus cytoplasmic masks are used to compute 35 cytological features in both channels. The difference of measurements between the nuclear and the nuclear plus cytoplasmic masks corresponds to a ring mask assigned as the cytoplasm. Hereafter, measures named as "cytoplasm" will be restricted to the cytoplasmic ring. Additionally, different segmentation approaches such as an autothresholding component are used to segment the image in the DAPI channel and define nuclear masks. This component segments images into foreground and background groups by applying *k*-means clustering with two clusters on pixel brightness. The new nuclear masks are used to compute 13 nuclear features in the DAPI channel that account for nuclear shrinkage. Image analyses were implemented within Pipeline Pilot platform taking advantage of its customizable components from image collection. Additional cytological features, such as nuclear perimeter, nuclear convex area, ..., or the same features but from different nuclear segmentation algorithms were considered and analyzed in order to maximize the information captured and annotated through this HCA. This additional information can be used in the future to identify other phenotypes of interest induced by the compounds, without the need for further analysis of crude images. In total, information for 48 cytological features (a table with this information is available as ESI†, Table S2) was captured and annotated in the DB together with the corresponding chemical information and assay conditions.

$$\text{Avg\_PCT\_PC} = \frac{\text{Avg\_Nuclear/Cytoplasmic\{well\}} - \text{Avg[Avg\_Nuclear/Cytoplasmic\{NegCtrl\}]}}{\text{Avg[Avg\_Nuclear/Cytoplasmic\{PosCtrl\}]} - \text{Avg[Avg\_Nuclear/Cytoplasmic\{NegCtrl\}]}} \times 100 \qquad (1)$$

Hit assessment

For the purpose of hit identification, and as in the case of the standard approach, the fluorescence intensities in the GFP channel image from the DNA and non-DNA regions were selected from the 48 annotated cellular features. Two parameters, nuclear cytoplasmic_difference and nuclear/cytoplasmic ratio, were computed for each cell and averaged over all cells in a well. Then, average values of these two parameters were referred to the negative/positive control wells to compute the phenotypic profile (Avg_PCT_PC and P70_Ratio-P50_PC values) and derive two metrics (Hit and Hit_R) that define the active criterion for the screening. Two parameters were derived for each cell: (i) nuclear/cytoplasmic ratio, which is the nuclear pixel intensity mean divided by the cytoplasmic ring pixel intensity mean and (ii) nuclear_cytoplasmic difference, which is the difference of the nuclear pixel intensity mean minus the cytoplasmic ring pixel intensity mean. Per each well, the average value of nuclear_cytoplasmic difference over all cells as well as the 70th-percentile of the nuclear/cytoplasmic ratio distribution were calculated. For control wells, the median (50th-percentile) of the nuclear/cytoplasmic ratio was considered. Each compound was assigned two flags (Hit and Hit_R) in order to be classified as hit or not. A well has Hit = 1 if the difference between its average nuclear_cytoplasmic difference and the mean of the negative controls is equal to or higher than the difference between the mean values of the positive and negative controls of the plate (eqn (1)); in this paper, these Avg_PCT_PC values are reported as percentages. For the calculations, only controls within the same plate as the test compound were considered; thus, during this automatic hit assessment process a dynamic threshold definition, plate-based, was considered to balance inter-plates experimental variability. Otherwise, Hit = 0 is assigned to the well. This Hit metric is quite restrictive for a primary screening, as a compound must produce a signal equal to or above 100% of the signal obtained for the average of the positive controls. Therefore, to minimize the effect of potential errors that encompass assay artifacts derived from high-content primary screening a complementary less restrictive hit definition, Hit_R, was applied too. Concerning Hit_R, a value of 1 was assigned to those wells having 70th-percentile of the nuclear/cytoplasmic ratio higher than the median of the positive control wells of the plate (P70_Ratio-P50_PC). In this way, we require for a well to be a hit if at least 30% of the cells present a nuclear accumulation of fluorescence similar to the ratio shown by 50% of cells in the positive controls. This criterion, P70_Ratio-P50_PC, was defined after an optimization process where a small set of compounds, 2.8% of the total (10 plates), was utilized to assess optimal percentile to maximize the number of identified hits according to assay conditions.

Although at this stage data analysis is implemented within the same routine that extracts information from images, all measurements were already automatically exported to raw text files containing information on a per cell basis.

Compounds with Hit = 1 and Hit_R = 1 were considered hits (positive), compounds with Hit = 0 and Hit_R = 0 were regarded as non-hits (negative) and compounds having either Hit or Hit_R equal to 1 were classed as "fuzzy" hits (Fig. 2A) which are flagged to be analyzed and categorized later by a Bayesian classifier.

The biological space derived from standard translocation cell-based assay is represented by its key phenotypic profile, hit/non-hit; therefore, by Avg_PCT_PC and P70_Ratio-P50_PC values. Thus, graphical representation of these statistical criteria utilized for hit assessment, which are based on data from annotated cytological features, provides a clear picture about assayed compounds in the biological space. The chemical space defined by these 33 992 screened structures is color-coded by hit assessment criteria defined in the biological space (Fig. 2B); but, this is just a qualitative overlapping

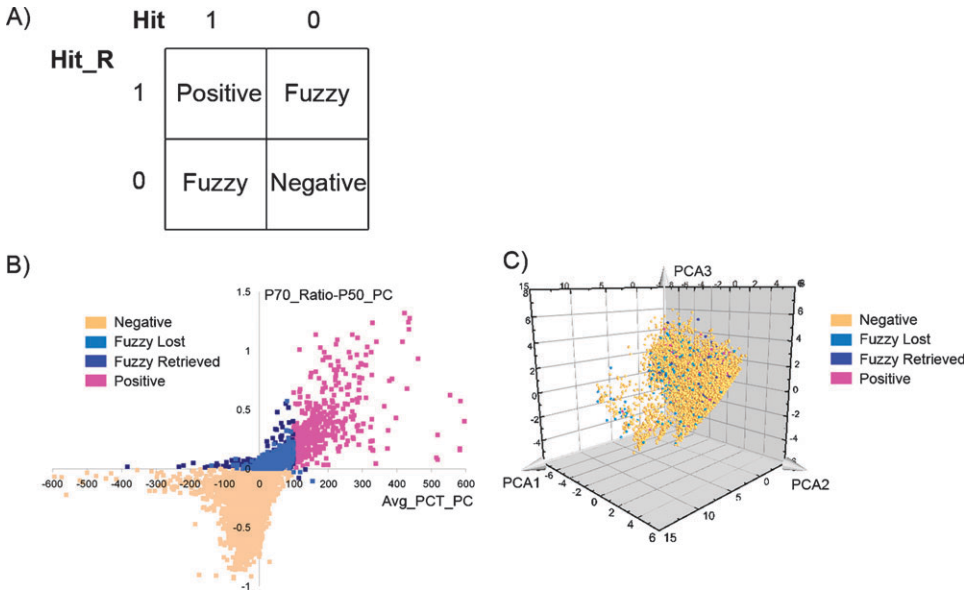

Fig. 2 (A) Confusion matrix showing criteria used to define hits, non-hits and fuzzy compounds. (B) Biological space is defined by key statistical criteria for phenotypic profiling (Avg_PCT_PC and P70_Ratio-P50_PC) and (C) graphical representation of the chemical space determined by three principal components (PCA1, PCA2 and PCA3), derived from ECFP_6 fingerprints,[18] which covers 31.5% of the chemical space defined by the 33 992 compounds.

between structural fingerprints (capitalized as Principal Components) and phenotypic response. Quantitative integration of both scenarios is necessary in order to maximize the quality of extracted information; thus, a Bayesian model based on biological and chemical spaces was built.

This statistical categorization method has been shown to be useful in creating models that analyze high-throughput screening data sets even those with noisy data;[18,19] therefore, together with the two hit criteria a Bayesian model (BM) based on the integration of chemical and phenotypic information was used as part of this automated HCS hit identification process with the objective to minimize, as much as possible, assay artifacts and, thus, maximize hit assessment (true positives).

A two-class Laplacian-modified naïve Bayesian model was trained with those compounds identified as hits and non-hits using as descriptors extended-connectivity fingerprints with a neighborhood size of six (ECFP_6)[18,20,21] (more details about BM are in the Experimental section). This additional step has a very low computational cost, scales linearly with the number of samples and does not require tuning parameters beyond the selection of the input descriptors. Hits and non-hits (Fig. 2A) compose a training set to generate a Bayesian model that is later applied to predict the relative likelihood of a compound from the "fuzzy" hit pool being a member of the hit subset. Additionally, this categorization model may be employed to categorize compounds for which image acquisition failed or image analysis was not possible (*e.g.*, because of precipitation of fluorescent dyes). Thus, Bayesian classifier is implemented in the workflow as an automated iterative feedback tuning to assess "fuzzy" hits and images.

Once compounds initially classified as "fuzzy" are assessed as hits or as non-hits by the model, the automated process is ready to finish after annotation of the DB. Now, each of the assayed compounds is annotated in the chemical-biology database with its 48 cytological features as well as with its phenotypic profile. All this information, ROI summary, is stored as text and for this set of 33 992 assayed compounds occupies 6 GigaBytes. The total time to process the image analysis of all 96-well plates containing these compounds was approximately 35 hours while the data analysis step (including Bayesian model and database annotation) was less than 2 hours (detailed information about DB annotation and server is provided in the Experimental section).

The reliability and feasibility of this one-step HCS analysis was assessed by comparing the compounds confirmed as hits, after secondary screening, using the BD Pathway Bioimager with the conventional data processing procedure (a table containing this information is available as ESI†, Table S3) with those classed as hits using the automated approach, Table 1. To illustrate and track all details about the hit definition process implemented in this system, those "fuzzy" compounds conclusively categorized by the Bayesian model as hits are reported as "fuzzy retrieved" and non-hits as "fuzzy lost". Some representative examples of 6 different chemotypes identified as hits are reported in Fig. 3.

Of the 539 (1.59% of full-library containing 33 992 small molecules) compounds re-tested in a second HCS round, 228 (42%) were confirmed as FOXO nuclear translocation inhibitors.[14] After running this one-shot HCA, using the 33 703 images from the first HCS (289 were not available), 521 compounds (1.5%) met the active criterion (Hit = 1 and Hit_R = 1) and 998 satisfied at least one metric (Hit or Hit_R) and were further classed as hits by the Bayesian model, "fuzzy retrieved" compounds. Of the 228 secondary confirmed hits 211 were properly identified by this new method, 177 (77.6%) were found within the 521 hits and 34 (14.9%) were in the 998-membered set recovered by the two-class Bayesian; then, 17 (7.5%) were misclassified. Thus, 92.5% of the confirmed

hits would have been identified if only this automated protocol were used for analyzing images from the first HCS campaign. Furthermore, upon checking the images from those wells labeled as hits by this new method, 40 additional primary hits that had been classified as negatives with the classical approach, BD analysis, were identified. 27 of them met Hit criteria, using inter-plate dynamic threshold values to identify hits makes the difference *versus* classical approach, as well as Hit_R definition. In addition, other 13 new hit compounds were definitively categorized as "fuzzy retrieved" by virtual screening: the Bayesian classifier.

Therefore, the automated method identified 92.5% of the hits (true positives) as well as 96.1% of the non-hits: 32 184 true negatives (30 150 non-hits and 2034 "fuzzy lost") and 1308 false positives (1519 compounds classed as hits but only 211 were real hits). Considering the diagonal from the confusion matrix, 211 true positives and 32 184 true negatives, this gave an overall accuracy of 96.1% (33 703 images were analyzed). The main drawback was the relatively high number of images labeled as hits but false positives, 3.8% of the total (1308 images). Interestingly 40 additional hits were identified enriching in a 14.9% the final number of hits obtained, 268. In addition, the positive impact of the Bayesian model in the process was readily apparent as 14.9% of the 228 hits were identified from those compounds initially classified as "fuzzy".

Compounds bearing the following chemotypes, described in Fig. 3, class A: phenotiazines (ETP-16151, ETP-16552, ETP-16182),[22] class B: hexahydrocarbazoles (ETP-16428,

Table 1 Results from the novel one-shot approach. Classification of the 33 992 membered library as well as details about the categorization of its corresponding 228 Hits identified using the classical approach *via* BD Pathway Bioimager[14]

	Compound set (33 992)	2nd hits (228)	Additional hits
Positives	521 (1.5%)	177 (77.6%)	27
Fuzzy retrieved	998 (3.0%)	34 (14.9%)	13
Fuzzy lost	2034 (6.0%)	16 (7.01%)	
Negatives	30 150 (88.7%)	1 (0.44%)	
No images available	289 (0.9%)		

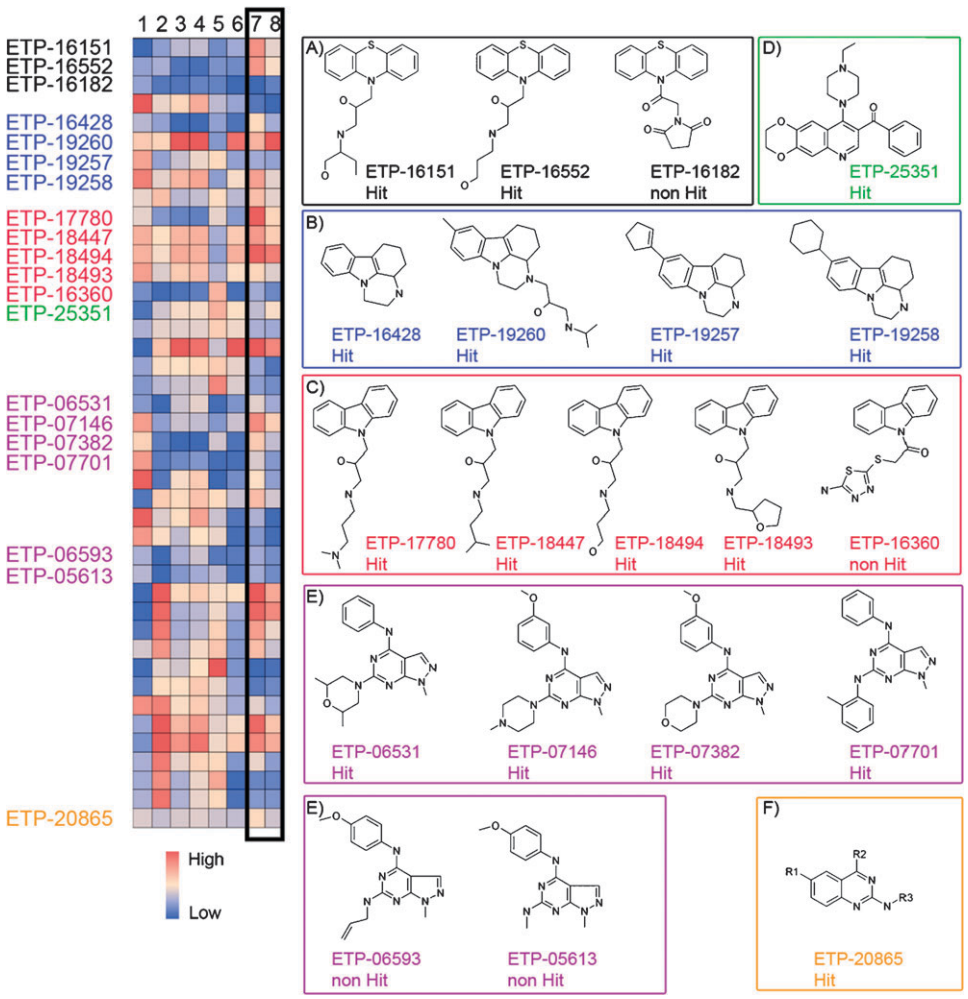

Fig. 3 Representative set of FOXO hits and non-hits, bearing 6 different chemotypes, (A–F) described in the main text, together with heat map (rows, compounds; columns, features) for selected set of cytological features [1: cell density; 2: nuclear DAPI intensity mean; 3: nuclear GFP intensity mean; 4: cytoplasmic GFP intensity mean; 5: nuclear area; 6: average nuclear_cytoplasmic difference] and derived statistical values driving to phenotypic assessment [7: Avg_PCT_PC; 8: P70_Ratio-P50_PC]. Then, just for columns 7 and 8, closer to red means higher probability to be labeled as hit (Hit and Hit_R respectively) and closer to blue as non-hit.

ETP-19260, ETP-19257, ETP-19258),[22] class C: carbazoles (ETP-17780, ETP-18447, ETP-18494, ETP-18493),[22] class D: quinoline-derived series analogs to quinostatin (ETP-25351)[23] and class E: pyrazolopyrimidines (ETP-06531, ETP-07146, ETP-07382, ETP-07701, ETP-06593 and ETP-05613),[14,24] have already been reported as PI3K inhibitors (hits and non-hits).

ETP-18493, ETP-19257 and ETP-19258 are examples of compounds overlooked using the standard approach but were identified by the automated one-step process, two by the Bayesian model and one as a positive hit satisfying two metrics (Hit and Hit_R). ETP-25351, ETP-16428, ETP-06531 and ETP-07701 exemplify cases recovered by the Bayesian model that are very closely similar analogs to those structures considered in the training set. However, among the 34 hits recovered by the Bayesian, two additional chemotypes were found that were not present in the training set, indicating that the Bayesian model not only recognizes similar compounds, but also identifies chemotypes different from those in the training set. One of these two chemotypes is exemplified by the scaffold borne on ETP-20865, class F: 2-aminoquinazoline, derivatives of which have also been recently reported as PI3K inhibitors.[25] Structural details for the additional chemotype are not shown.

Post-process data analysis—identification of additional phenotypes: cytotoxicity

This automated *one-step* HCA allows one to exploit the fact that HCS, unlike traditional screening methods, can simultaneously collect a variety of compound-induced phenotypic information in addition to that used for hit identification including the 48 cytological features described in Table S2 (available as ESI†). For the screen described above these data together with the corresponding processed information and phenotypic profile were directly deposited in the Oracle-based relational database for each assayed compound. All this information, ROI summary, is archived as text together with a link to original images, which is located in an independent repository, and can be accessed *on-the-fly* if it is requested (Fig. 4). Therefore, data obtained and derived from HCS are accessible at any time by all users *via* a user-friendly web application. All this is implemented within Pipeline Pilot platform allowing an efficient integration with an Oracle database, in which captured and analyzed information is deposited, and with the server where crude images were uploaded.

To test whether the analysis of the annotated information extracted from image analysis can provide information about a compound's potential cytotoxicity we screened 168 additional compounds including 12 cytotoxic reference compounds (*e.g.* gliotoxin, Fig. 5B) under identical conditions to the PI3K/Akt screen described above. Then we performed multivariate statistical analysis of the data in order to understand which phenotypic features provide a proper assessment of cytotoxicity. Cytotoxic compounds cause cells to exhibit morphological hallmarks of apoptosis, such as DNA fragmentation, nuclear condensation and segmentation. Moreover, a notable decrease in cell density is found.[26,27]

Statistical comparisons among the 12 cytotoxic compounds and the negative controls (untreated wells) and among the 156 non-cytotoxic compounds (plus positive controls) and the negative controls were conducted with Student's *t*-test in order to identify those key cytological features (explicit descriptions about performed Student's *t*-test and assessment criterion, including details about kappa coefficient,[28] are provided in the Experimental section). This analysis suggested that the main features to identify a cytotoxic phenotype are: cell density, large variations in DAPI intensity along the nuclear mask (increased pixel intensity variance) and decreased morphological properties of the nuclei (area, perimeter, ...).

Based on the 7 key properties with significant *p*-values in Student's *t*-test (Table S4 and Fig. S1, available in the ESI†), these data were used as a training set to implement criterion to identify compounds that induce a cytotoxic phenotypic profile.

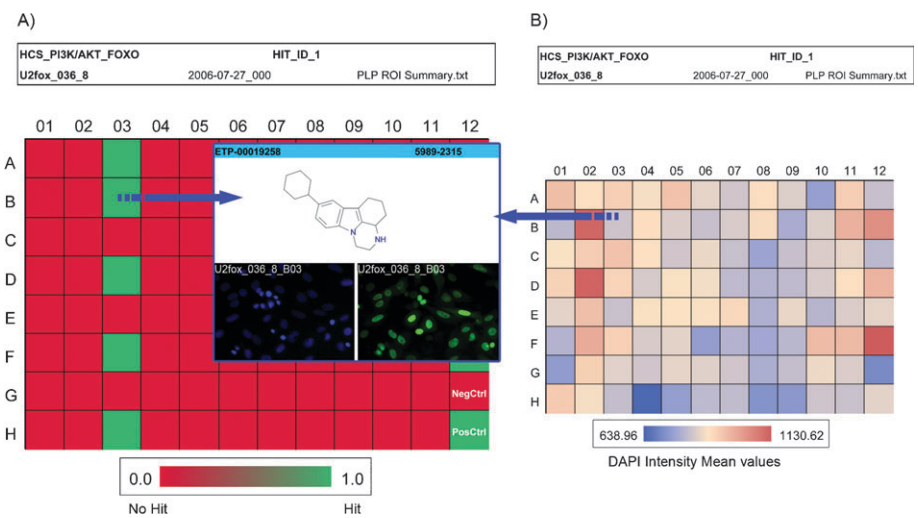

Fig. 4 Plate-based heat map information. (A) In this example, Hit ID flag is reported (hit = 1, and non-hit = 0) and values (variance, mean, ...) for any annotated cytological features (DAPI_intensity, nuclear area, ...) can be also plotted and reported; in case (B) heat map shows mean for DAPI intensity values. In addition, this dynamically generated document is linked to corresponding images and chemical information; therefore, by double-click any well of interest this information is retrieved too—as it is represented in these two cases (A) and (B).

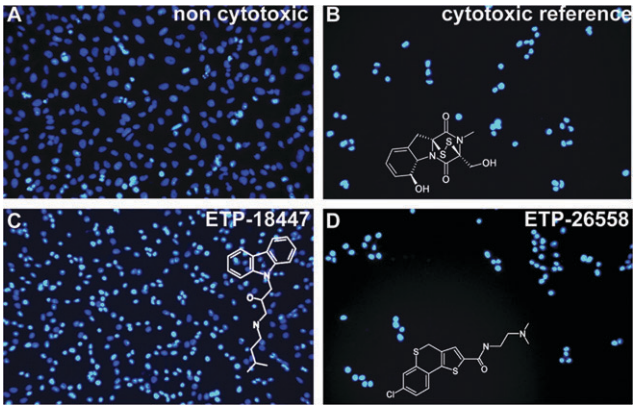

Fig. 5 DAPI images of U2foxRELOC cells following treatment with cytotoxic agents. U2foxRELOC cells were seeded in 96-well plates, incubated for 12 h and treated with DMSO (A) or gliotoxin (B) which was used as a reference cytotoxic compound. Incubation of U2foxRELOC cells in the presence of the library compounds ETP-18447, FOXO hit, (C) and ETP-26558, FOXO non-hit, (D) resulted in cytotoxicity of varying degrees. In fact, ETP-26558 has been very recently reported as cytotoxic compound.[29]

Thus, by comparing different kappa values[28] and with the aim of allowing for flexibility in the definition, the following rule was finally established: a well will be classified as cytotoxic if at least six out of the seven significant cytological features are outliers; *i.e.*, their mean value is two standard deviations away from the mean values of the negative controls. This flexibility drove to the identification of cytotoxic compounds such as ETP-18447 (Fig. 5C) where cellular density was right but the rest of key cytological features met the requirements, in this case nuclear condensation is an important feature, or ETP-26558 (Fig. 5D) where cellular density is clearly deficient. This criterion was ruled out by maximizing agreement in the confusion matrix (table with this information is available as ESI†, Table S5).

Thorough visual inspections were conducted on the original images of 10 randomly selected plates (874 compounds) to identify unambiguous cytotoxic activities. On analysis of the annotated HCS DB using the previously defined criterion we found that 13 out of 14 unambiguously cytotoxic compounds of the validation set were properly labeled as cytotoxic compounds (92.8%—true positives) and 838, out of 860, were identified as non-cytotoxic compounds (97.4%—as true negatives).

Therefore, at least in this case, we can conclude that analysis of the annotated cytological features can reliably identify compounds with different phenotypic profile to the initial aim of the HCS. Thus, this automated *one-step* process that archives all the information from HCA provides additional value since it allows querying of new phenotypic profiles, at any time without the need to re-analyze the initial crude images.

Discussion

This manuscript describes the development, implementation and validation of a customizable automated tool which requires no user intervention to analyze initial crude images from a HCS campaign in one step. The novel system automatically: (a) extracts and measures all cytological features requested, (b) performs statistical analysis, including Bayesian model integrating phenotype profiling and chemical space, (c) performs Hit assessment. Based on the targeted phenotype, (d) archives all information captured through this HCA together with chemical structural data in an annotated Oracle-based relational database. This fully automated *one-step* process, when used to analyze images from a screen of 33 992 compounds, had an overall accuracy of 96.1% with an optimal balance for sensitivity and specificity of 1.04; in addition, the automated system identified 40 additional hits that were lost in the conventional user-dependent multistep process. However, relatively a large number of false positives was obtained, 3.8%. And, (e) post-process data analysis: identification of additional phenotypes, cytotoxicity. Once the large amount of data extracted from HCA is annotated in a relational database, a user-friendly access for data retrieval and visualization tools is key since they are critical to re-analyze cytological features that may be highlighted as important at a later time. Thus, a web-based interface was implemented to facilitate access to annotated information as well as to crude images. In fact, multivariate statistical analysis of previously annotated features facilitated the identification of key feature to best assess cytotoxic compounds. In this case, in which 874 compounds

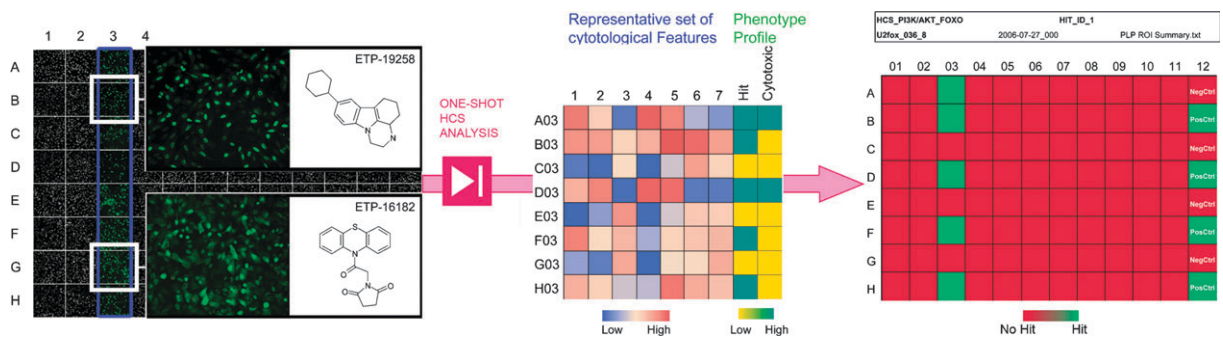

Fig. 6 From images to annotated DB: cytological features as well as phenotype profile(s) are extracted, analyzed and annotated in just one step without user intervention. In addition, user-friendly interface was implemented to retrieve and analyze all stored information. Cytological features in heat map correspond to [1: Avg_PCT_PC; 2: P70_Ratio-P50_PC; 3: nuclear area; 4: nuclear GFP intensity mean; 5: nuclear DAPI intensity mean; 6: cell density; 7: nuclear equivalent diameter], where values have been scaled between [−1.5 and 1.5] for representation purposes. Phenotypic profile(s) and cytotoxicity classification are obtained from each image.

were evaluated, an overall accuracy of 97.3% was achieved and the balance for sensitivity and specificity was very good at 1.05.

To the best of our knowledge, this is the first reported "*non-stop trip*" from images to annotated DB. The tool has been developed and validated against a relatively large HCS assay, 33 992 compounds, yielding good results: not only speeding the drug discovery process up but also adding important quantitative and qualitative values, in terms of data extraction, annotation, analysis and hit identification. This system can be customized to different image-based phenotypic screening campaigns. In fact, minor adjustments to the current system are required if analyzed HCS campaigns are focused on translocation assays or the aim is the identification of cytotoxic compounds. But, a different experimental set up, *i.e.*, including dyes for cytoplasm location, would enforce a different image analysis procedure which may be the most difficult step for non-experts. For the rest of the procedure, data analysis (in case different metrics were required), reporting results and database annotation, the implemented method is quite general; thus, specific modifications can be easily adapted. Fig. 6 summarizes the whole one-step process, implemented within a Pipeline Pilot platform (Fig. S3, as ESI†), described in the paper.

Experimental section

Compounds and data collection

Compounds were acquired from ChemDiv (San Diego, CA), BioFocus (Cambridge, UK) and the Biomar Institute (Leon, Spain). In all cases, compounds are at least 90% pure substances with structure and purity validated by NMR and/or LCMS. Detailed protocols for the preparation of the samples and images for the 33 992-compound compendium, and for the additional set of 168 compounds, are as previously described.[13]

High-content screening assay

The cloned cell line, U2FoxReloc, was isolated after transfection of a GFP–FOXO3a reporter plasmid into U2OS cells and responds to inhibition of different signaling events that regulate the nuclear–cytoplasmic shuttling of FOXO transcription factors: PI3K/Akt signaling, nuclear export and calcium/CaM-dependent signaling. Moreover, known ATP-competitive PI3K inhibitors with different potencies (LY294002, wortmannin, PIK-75, and PI-103) respond in a dose-dependent manner, agreeing with IC50 values reported for PI3Kα inhibition *in vitro*. This HCS was utilized to primary screen a library of 33 992 compounds which were administered at a single dose of 10 μM, maximal concentration of DMSO was 1%, and incubated for 1 h at 37 °C. Each compound was screened in singles. The nuclear export inhibitor LY294002 served as a positive control and "landmark" for the phenotype likely to be induced: *i.e.*, the nuclear accumulation of fluorescence triggered by LY294002 was defined as 100% activity while untreated wells (only DSMO) were the negative controls. GFP functional interference was discarded after being checked. The BD Pathway Bioimager was used to automatically capture images (TIF format) using a 10× dry objective.[13]

Image and data analysis *via* BD Pathway Bioimager

All the initial image and data analysis was performed by using BD Pathway Bioimager through the AttoVision software and the Image Data Explorer for data processing. For image analysis, the automated segmentation algorithm (Auto-ROI) with ring (two outputs) option was chosen to create two ring shaped regions of interest (ROIs) around the DAPI dye. The inner ring defines the nuclear mask, whereas the outer, dilated 1 pixel out from the other, labels the cytoplasmic region. The nuclear/cytoplasmic (Nuc/Cyt) ratios of fluorescence intensity were determined by dividing the intensity of the GFP fluorescence from the nucleus by that in the cytoplasm. Nuclear accumulation of fluorescent signal for each cell was defined by applying a fixed threshold ratio of greater than 1.8.[13] Based on this procedure we calculated the percentage of cells per well exhibiting nuclear translocation. Compounds that induced nuclear accumulation of the fluorescent reporter above 60% of the signal obtained from wells treated with 20 μM LY294002 were considered as hits.

Descriptors and Bayesian classifier

The method used in this work is based on circular molecular fingerprints using extended connectivity fingerprints (ECFPs) descriptors,[18,20] implemented in Pipeline Pilot. Laplacian-modified Bayesian[18,30] analysis was evolved from naïve Bayesian analysis to deal with the high-dimensional representation of molecules using extended-connectivity fingerprint. ECFP descriptors define molecular structure using radial atom neighborhoods. In particular, ECFP of size 6 (ECFP_6) contains all circular substructures around each atom, up to a maximum width of 6 bonds, capturing explicit atom-type information for each atom.[18] The total set of patches characterizes the whole molecule.

The predictive naïve Bayesian model was created from the trained data (hits and non-hits), and was firstly validated using 4-fold cross-validation: one fourth of the samples were left out and the model was built using the remaining samples; that model was used to predict the scores for the left-out samples. After sorting compounds by decreasing score, a receiver operating characteristic (ROC) plot was built to estimate the accuracy of the model, yielding a ROC score of 0.96.

Post-process data analysis: cytotoxicity assessment

Student's *t*-test. Two-tailed *t*-tests were applied for 13 out of the 48 collected cytological features (average well's values) plus the cell density measurement. The average and standard deviation of the ratio values of the negative controls were calculated across all wells within the same plate, in order to surpass inter-plate variability. Table S4, available as ESI†, summarizes *p*-values found for each of the three 96-well plates. A significance level of 1% or better ($P < 0.01$) was required for a cytological feature to be included in the posterior analysis. As *p*-values vary across plates, a consensus on the cytological features was required resulting in a total of seven finally selected cytological features (in bold in Table S4, as

ESI†) that satisfy significance level for all three plates. In Fig. S1, available as ESI†, the well's average value for each of the seven cytological features is depicted for cytotoxic compounds (blue dots) and non-cytotoxic and positive/negative controls (black crosses). The effect of a toxic compound in the cytological feature can be easily recognized: reduction in equivalent diameter, area, perimeter, area of the convex hull and cell density and increased mean pixel intensity and intensity variance (accounting for big changes in pixel intensity across the nuclear region).

After identifying key cytological features, different criteria to define "cytotoxic" compounds were analyzed in order to maximize agreement in the confusion matrix (kappa statistics) for the 168-membered reference library. Independent models were created, where each feature was separately considered to be "outlier" if the wells' average value was 1, 1.5, 2 or 3 standard deviations away, respectively, from the average response of the negative control wells. For each independent model, a minimum number of "outlier" features out of the seven was required in each case to define a compound as toxic. Using these definitions, all compounds from the library were classified and the corresponding confusion matrix with sensitivity, specificity and kappa values was calculated (Table S5, available as ESI†). For the 168-membered library, training set, the selected criterion to identify cytotoxic phenotypic profile delivered a prediction of 100% of cytotoxic compounds (100% sensitivity) with reasonable specificity (97%); but, these are not the best results described in Table S5—aim, avoid overfitting.

Kappa coefficient. The principle of kappa index (κ) is to measure the difference between the agreement of the working process and the agreement of chance in the classification process. Kappa statistics can be interpreted as a measure of agreement that exists beyond the degree expected by chance alone. This coefficient makes full use of the information contained in the confusion matrix, not only the diagonalisation (overall accuracy) but also those entries containing the number of false positives and false negatives. The kappa coefficient is the optimal measurement to determine if a classifier predicts better than chance.[28] Kappa has a range of values that typically range from 0 to 1, with 1 indicating perfect agreement and 0 indicating a pattern arising by chance. It is calculated from the analysis of the confusion matrix (Fig. S2, available as ESI†) and it is of the form (eqn (2)):

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

where P_o is the observed probability or observed accuracy (the matrix diagonal) (eqn (3)) and P_e is the expected probability, the expected proportion of matches in this diagonal assuming a model of classification independence derived from the observed row and column totals (eqn (4)).

$$\text{Observed probability } P_o = \frac{TP + TN}{N} \quad (3)$$

$$\text{Expected probability } P_e = \frac{\left[\frac{TP+FN}{N} + \frac{FP+TN}{N}\right]}{N} \quad (4)$$

Sensitivity (eqn (5)) that measures the proportion of actual positives which are correctly identified, and specificity (eqn (6)), that measures the proportion of negatives which are correctly identified, were also calculated.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100 \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100 \quad (6)$$

Database annotation

Standard Open Database Connectivity-compliant (ODBC) components available in Pipeline Pilot are used to insert and update with SQL functionality our Common Chemical and Biological Repository (CCBR).

Server characteristics

HP Proliant DL 380 G4. 8 GB RAM. 2 CPU Intel Xeon 3.4 GHz with Hyperthreading disabled.

Abbreviations

BM	Bayesian model
CCBR	Common chemical and biological repository
Ch	Channel
DAPI	4′,6′-Diamidino-2-phenylindole
DB	Database
DMSO	Dimethyl sulfoxide
ECFP	Extended-connectivity fingerprint
FOXO	Forkhead box sub-group "O"
GFP	Green fluorescent protein
HCA	High-content analysis
HCS	High-content screening
IC50	50% Inhibitory concentration
ODBC	Open database connectivity-compliant
PCA	Principal component analysis
PI3K	Phosphoinositide 3-kinase
ROC	Receiver operating characteristic
ROI	Region of interest
TIF	Tagged image file

Acknowledgements

We thank L. Bleicher, T. Moran and I. Mikic from Accelrys for their assistance and help with "image collection" within Pipeline Pilot as well as for their stimulating discussions and suggestions, and M. Urbano-Cuadrado from Experimental Therapeutics Programme for his assistance and help in database annotation, and A. Fernandez from the Biomar Institute for providing a library of 168 compounds, and O. Rueda and R. Diaz for their assistance with statistical analyses. This work was supported by funding from the Spanish Ministerio de Ciencia e Innovacion (Project BIO2006-02432).

References

1 K. A. Giuliano, R. L. DeBiasio, R. T. Dunlay, A. Gough, J. M. Volosky, J. Zock, G. N. Pavlakis and D. L. Taylor, *J. Biomol. Screening*, 1997, **2**, 249–259.
2 A. E. Carpenter, *Nat. Chem. Biol.*, 2007, **3**, 461–465.

3 U. S. Eggert and T. J. Mitchison, *Curr. Opin. Chem. Biol.*, 2006, **10**, 232–237.
4 T. J. Mitchison, *ChemBioChem*, 2005, **6**, 33–39.
5 R. A. Blake, *Methods Mol. Biol.*, 2007, **356**, 367–377.
6 P. Lang, K. Yeow, A. Nichols and A. Scheer, *Nat. Rev. Drug Discovery*, 2006, **5**, 343–356.
7 C. Smith and M. Eisenstein, *Nat. Methods*, 2005, **2**, 547–555.
8 S. Lee and B. J. Howell, *Methods Enzymol.*, 2006, **414**, 468–483.
9 A. E. Carpenter, *Nat. Methods*, 2007, **4**, 120–121.
10 K. A. Giuliano, J. R. Haskins and D. L. Taylor, *Assay Drug Dev. Technol.*, 2003, **1**, 565–577.
11 T. R. Jones, A. E. Carpenter, M. R. Lamprecht, J. Moffat, S. J. Silver, J. K. Grenier, A. B. Castoreno, U. S. Eggert, D. E. Root, P. Golland and D. M. Sabatini, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 1826–1831.
12 A. Di Cristofano and P. P. Pandolfi, *Cell*, 2000, **100**, 387–390.
13 F. Zanella, A. Rosado, B. Garcia, A. Carnero and W. Link, *ChemBioChem*, 2008, **9**, 2229–2237.
14 W. Link, J. Oyarzabal, B. G. Serelde, M. I. Albarran, O. Rabal, A. Cebriá, P. Alfonso, J. Fominaya, O. Renner, S. Peregrino, D. Soilán, P. A. Ceballos, A. I. Hernández, M. Lorenzo, P. Pevarello, T. Gonzalez-Granda, G. Kurz, J. R. Bischoff and A. Carnero, *J. Biol. Chem.*, 2009, **284**, 28392–28400.
15 Pipeline Pilot v5.1 available from Accelrys: http://accelrys.com/. Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA.
16 F. Meyer, PhD thesis, Ecole des Mines, 1979.
17 N. Otsu, *IEEE Trans. Syst. Man Cybern.*, 1979, **9**, 62–66.
18 D. Rogers, R. D. Brown and M. Hahn, *J. Biomol. Screening*, 2005, **10**, 682–686.
19 M. Glick, A. E. Klon, P. Acklin and J. W. Davies, *J. Biomol. Screening*, 2004, **9**, 32–36.
20 J. Hert, P. Willet, D. J. Wilton, P. Acklin, K. Azzoui, E. Jaboby and A. Schuffenhauer, *Org. Biomol. Chem.*, 2004, **2**, 3256–3266.
21 A. Bender, H. Y. Mussa and R. C. Glen, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 170–178.
22 T. R. Kau, F. Schroeder, S. Ramaswamy, C. L. Wojciechowski, J. J. Zhao, T. M. Roberts, J. Clardy, W. R. Sellers and P. A. Silver, *Cancer Cell*, 2003, **4**, 463–476.
23 J. Yang, A. Shamji, S. Matchacheep and S. L. Schreiber, *Chem. Biol.*, 2007, **14**, 371–377.
24 A. Zask, P. W. Nowak, J. Verheijen, K. J. Curran, J. Kaplan, D. Malwitz, M. G. Bursavich, D. C. Cole, S. Ayral-Kaloustian, K. Yu, D. J. Richard and M. Lefever, *WO Pat.*, 115 974 A2, 2008.
25 N. D. Adams, J. L. Burgess, M. G. Darcy, S. D. Knight, K. A. Newlander, L. H. Ridgers and S. J. Schmidt, *WO Pat.*, 157 191 A2, 2008.
26 S. B. Tencza and M. A. Sipe, *J. Appl. Toxicol.*, 2004, **24**, 371–377.
27 V. C. Abraham, D. L. Towne, J. F. Waring, U. Warrior and D. J. Burns, *J. Biomol. Screening*, 2008, **13**, 527–537.
28 A. H. Fielding and J. F. Bell, *Environ. Conserv.*, 1997, **24**, 38–49.
29 M. Ciustea, J. E. Silverman, A. M. Druck-Shudofsky and R. P. Ricciardi, *J. Med. Chem.*, 2008, **51**, 6563–6570.
30 X. Xia, E. G. Maliski, P. Gallant and D. Rogers, *J. Med. Chem.*, 2004, **47**, 4463–4470.

