# Visualization and Interpretation of High Content Screening Data

Andrew Smellie,*,† Christopher J. Wilson,‡ and Shi Chung Ng‡

Departments of Informatics and Modeling and Chemical Genomics, ArQule Incorporated, 19 Presidential Way, Woburn, Massachusetts 01801

Received September 15, 2005

High content screening is a method for identifying small molecule modulators of mammalian cell biology. The nature of the experiment generates an enormous amount of data in the form of photographic images of cells after treatment with compounds of interest. The interpretation of data from these experiments is challenging both in terms of automatically perceiving the images, extracting, and understanding differences between screened compounds and visualizing the results. This paper discusses the application of statistical and visual methods that have been used to interpret data from a simplified DNA stain (DAPI) screen to quickly identify compounds of interest. An understanding of the mechanism of action of the screened compounds can be obtained by comparing them to control compounds of known mechanism of action. Statistical and visual methods will be shown that facilitate easy comparison of screened compounds against these control compounds. As an example, a subset of the internal repository at ArQule was screened, together with control compounds that were known to induce characteristic mitotic arrest. Subsequent data processing described in this paper permitted the easy identification of compounds that were similar to (and very different from) the control compounds.

## INTRODUCTION

High content screening based on microscopic images is a well-recognized platform for testing compounds in "phenotypic" cell-based assays.[1−4] In these assays fluorescently labeled mammalian cells are photographed, and the images are processed to identify cells and other cellular information at the molecular level. By a choice of suitable cell stain, or other fluorescent label, examination of the images can yield information on, for example, mitotic index changes,[5] the identification of wound-healing inhibitors,[6,7] and measurement of GPCR recycling.[8] Well-to-well (i.e. compound-to-compound) comparisons of the images allow unknown (i.e. screening) compounds to be compared against known standards. This paper discusses how efficient comparisons can be made between wells in a high throughput screen using a combination of a simple statistical measure coupled with low-dimensional projection techniques to fully characterize a compound and understand how it relates to one or more control compounds.

To identify compounds of interest that induce mitotic arrest, a high content assay was developed based upon chromatin staining. Induction of mitotic agents ultimately leads to cell death and has been particularly useful in treating late stage cancers. Both Taxol and Vinblastine are widely prescribed oncology drugs, and agents that have similar cellular affects are sought after. Here, a chromatin condensation[9] assay is used as a proxy for mitosis. Compounds to be screened were added to cells and incubated for 18 h in black-walled, clear-bottomed wells. The cells were stained with DAPI (Invitrogen DS21490), a DNA binding dye that is fluorescent only when bound to DNA. Sets of 4 images per well using automated microscopy were taken that comprised about 10% of the surface area of the well. The raw images were used to identify the cell bodies using software proprietary to the instrument (Beckman-Coulter Q3DM). Summary parameters for the individual cell bodies were computed using the pixel information from the images. A simple test based on the Kolmogorov-Smirnoff (KMG) statistic that compares distributions of these summary parameters was used to identify wells that cause an increase of cells in mitosis. The KMG effectively defines a "distance" between two wells. Trivial comparisons can then be made to see if a test compound is "close" to a control compound. Going a step further, dimensional projection techniques can be used to project these wells (in their high-dimensional KMG distance space) into a low dimensional coordinate space, such that the distances measured in 2D reproduce the true KMG distances as closely as possible. This method is often superior to automated clustering[10,11] because these clustering methods must make assumptions about which compound is in which cluster. The issue of the ambiguity of cluster membership is an area of ongoing research.[12] By projecting into a visualizable space, the choice of cluster membership can be left to the observer, who is often in the best position to judge these subjective cluster assignments. It will also be shown that visualization can naturally discover situations where the assay has performed unreliably and thus serve as a form of quality assurance.

## EXPERIMENTAL DETAILS

Full experimental details of the experiments have been given elsewhere,[5] but a short summary will be given here.

**Cells and Cell Culture.** A549 cancer cells were purchased from the ATCC and maintained in a 37 °C, 5% CO2 incubator. All cells were adherent cells, subcultured using

* Corresponding author e-mail: asmellie@arqule.com.
† Department of Informatics and Modeling.
‡ Department of Chemical Genomics.

10.1021/ci050404g CCC: $33.50 
Published on Web 12/21/2005

trypsinization, and grown in DMEM media (4.5 g/L glucose) supplemented with 15% heat-inactivated FBS, 10 mM L-glutamine, and 10 mM Hepes.

**High-Content Assays.** A549 cells were seeded in black walled, clear bottomed 96 well plates and incubated for 16−24 h. Test compounds diluted in DMSO, either from library screening plates or serial dilution plates, were further diluted in complete media and then added to cells; the overall dilution in media was 300-fold, and the final DMSO concentration was 0.33%. Eighteen hours post compound addition, a 1:1 mixture of Mirsky's fixative parts A (10×) and B (10×) (National Diagnostics HS-102) was added directly to the culture medium; resulting in a 1× final concentration of fixative. Ninety minutes later the plates were washed into phosphate buffer saline (PBS). For screening, DAPI (Invitrogen DS21490) stain was added to the plates to a final concentration of 1 $\mu$g/mL, and then each well was overlaid with mineral oil to prevent evaporation.

**Imaging and Image Analysis.** Imaging and image analysis was carried using a Beckman-Coulter/Q3DM EIDAQ100 instrument. Four images per well were captured, and DAPI channel images were taken with a 16.6 ms/0% gain exposure. Automated cell identification and measurement extraction were also carried out within the software using standard settings.

**Compound Collection.** A collection of 13,399 small molecules was assembled from ArQule's compound repository of over a 1,000,000 arrayable compounds. This library is one of the key components within our initiative to assemble a high quality screening repository. Compounds within this set are selected to be

- Leadlike
- High quality pure single compound per well
- Biological target independent
- Diverse representation of in-house repository
- Compounds able to support rapid hit-to-lead activities

Compounds were selected based on leadlikeness filters (molecular weight <= 450; ClogP[13] −3.5 to 4.5; LogS[14] >= −6 ($\mu$m); #rings <= 4; HB donors <= 5; HB acceptors <= 8; #rotatable bonds <= 7) and also diversity (from visual inspection) and availability across our source library. Using ArQule's high-throughput chemistry platform, all stores of selected compounds were purified by preparative HPLC chromatography, dried, reweighed, dissolved in DMSO, and analyzed for final purity by HPLC/MS.[15,16] This library was dispensed into single use 96 well plates, stored frozen at −20 °C, and used for all subsequent screening activities. All compounds were >80% pure judged by UV214 nm and ELSD, and the major product contained a single mass species of the predicted molecular weight. Spot checks and follow up work indicate that purity and concentration of nearly all compounds are maintained using our storage methods (data not shown).

## METHODS

**An Image-Based Assay for Mitotic Arrest.** To assay small molecule compounds for mitotic arrest we chose to use the change in DNA condensation states that cells undergo while progressing through the cell cycle. As a cell goes from G1 through S then G2, its DNA typically has a diffuse nuclear stain. Upon entrance into mitosis, a cell's DNA condenses and organizes itself into chromosome bodies that are necessary to equally separate the DNA into the resulting daughter cells. This phenomenon can be visualized using a fluorescent DNA stain, such as DAPI (4′-6-diamidino-2-phenylindole). The condensation of nuclei is visually striking and can be quantitatively measured on a cell-by-cell basis using digital images and cell finding and integration software. Within a population of asynchronously growing cancer cells in culture, a small percentage of the cells (approximately 5% for A549 cells) will be in mitosis at any given time. However, when a population of cells is treated with a mitotic arresting agent, such as 250 nM Taxol for 18 h, many of the cells (approximately 50% in this case) have an obvious condensed chromatin and the other 50% appear to be not in M phase or upon closer inspection appear to have fragmented nuclei. This is typical of Taxol treatment and in our findings, many mitotic arresting agents, although the percentages of cells affected vary depending on the agent. Because of these subtle yet diverse responses, we developed a sensitive statistical test to compare cell *populations* to untreated cells. Ultimately, we wanted an assay that would be generally applicable and one that would return one value per well so that we could rank compounds from screening campaigns.

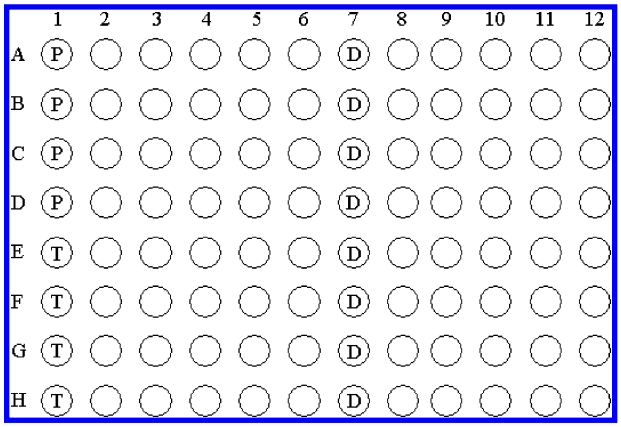


**Figure 1.** Standard plate layout used in the screening where: $P$ = first positive control compound at 10 $\mu$m (proprietary Arqule compound with known activity); $T$ = second positive control compound at 250 nm (Taxol); $D$ = negative control wells with only DMSO at a dilution of 0.33%. The rest of the wells contained screening compounds; thus each plate consists of positive and negative controls and test compounds.

**Data Analysis of Well Populations.** The layout of a typical plate during the screening is illustrated in Figure 1. Each 96 well plate consisted of positive control compounds in the form of Taxol (in wells E1-H1) and an internal compound (ArQ) known to induce cell mitosis (in wells A1-D1). The negative controls were wells with only DMSO (0.33% in wells A7-H7) at the same dilution used in all the other wells. The remaining wells (A2-H6, A8-H12) contained screening compounds in a 0.33% solution of DMSO. After high-content image screening, as described above, each well was photographed using automated microscopy to generate 4 images per well. Examples of these images are shown in Figure 2. There will be more explanation of these images in the subsequent results section.

Software proprietary to the data collection[17] instrument automatically perceived the locations of each cell in the image(s) and computed parameters for each cell. The list of computed parameters is given in Table 1. For each well there

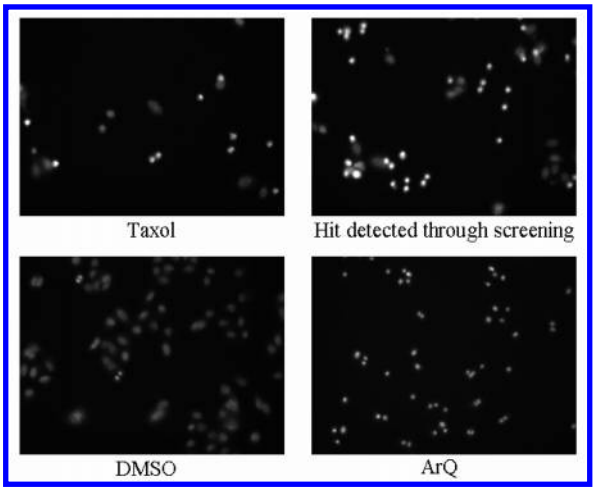


**Figure 2.** Example raw well images from automated microscopy.

**Table 1.** Parameters Computed Per Cell by the Beckman-Coulter/ Q3DM EIDAQ100

| parameter | definition |
|---|---|
| X | X position of the cell in the well (pixels) |
| Y | Y position of the cell in the well (pixels) |
| X_size | length of *X*-axis of the cell (pixels) |
| Y_size | length of *Y*-axis of the cell (pixels) |
| Width | X_size of the nucleus (microns) |
| Height | Y_size of the nucleus (microns) |
| Area | area of the nucleus (microns) |
| Area**1/2 | square root of the area |
| Perimeter | perimeter of the nucleus (microns) |
| Wiggle | perimeter/area |
| Wiggle_Nrm | perimeter /area**1/2 |
| Integ_GS | integrated gray-scale $\sum_{p \in \text{nucleus}} I(p).\text{area}(p)$ where $I(p)$ = pixel brightness normalized to 1 |
| Aver_GS | average gray-scale inside the nucleus normalized to 1 |
| Percent_0 | brightness of the pixels that is brighter than 0% of the nuclear pixels |
| Percent_5 | brightness of the pixels that is brighter than 5% of the nuclear pixels |
| Percent_25 | brightness of the pixels that is brighter than 25% of the nuclear pixels |
| Percent_50 | brightness of the pixels that is brighter than 50% of the nuclear pixels |
| Percent_75 | brightness of the pixels that is brighter than 75% of the nuclear pixels |
| Percent_95 | brightness of the pixels that is brighter than 95% of the nuclear pixels |
| Percent_100 | brightness of the pixels that is brighter than 100% of the nuclear pixels |
| IQ_Range | interquartile range (difference between the 25th and 75th percentile) |
| Variance | variance of nuclear pixel intensity |
| StdDev | standard deviation of nuclear pixel intensity |
| AbsDev | absolute deviation of nuclear pixel intensity |

were a number of computable parameters, and values were computed for each parameter for *every* cell in a well. In other words, each well can be characterized by a number of *distributions* of parameter values, one distribution per parameter.

The goal of the screening was to identify compounds that induced mitosis in cells upon treatment. Since the cell nuclei are stained with DAPI, and it is known that cell nuclei undergo characteristic physical and visual changes during mitosis, comparing the images of cells treated with test compounds with cells treated with DMSO (i.e. the negative controls) would indicate potential new mitotic-inducing compounds. This comparison was done by comparing the distribution of parameter values derived from the images for a test well with those of a control well. A well-established method of comparing population distributions is the Kolmogorov-Smirnoff[18] statistic (KMG), defined as the maximum difference in the fraction of values present in each cumulative distribution over all values $x$ in the distribution. A KMG value of 0 means the distributions are identical, and a value of 1 means the distributions are nonoverlapping and distinct. The KMG defines a "distance" between two distributions.

The actual score used to measure the difference between a test well $W$ and the negative control DMSO wells $\{D\}$ is given by eq 1 (the corrected Kolmogorov score used to assess compounds)

$$S(W,D,P) = \left( \frac{\sum_{i=1}^{N_D} \text{KMG}_P(W,D_i)}{N_D} \right) - B_C \qquad (1)$$

where $S(W,D,P)$ = the corrected Kolmogorov of the $W$th test well, when compared against the set of negative controls wells $D$ using parameter $P$, $W$ = the current test well being compared, $\text{KMG}_P(W,D_i)$ = the Kolmogorov-Smirnoff statistic computed using the distribution of parameter $P$ values in the test well $W$ and the $i$th control well $D_i$, $P$ = the choice of parameter used for KSS computation (e.g. average gray scale, AVER_GS), $D_i$ = the $i$th negative control well on the plate containing only DMSO, $N_D$ = the number of control wells (8) per plate, and $B_C$ = the "background" KMG value, defined as the mean KMG distance between all pairs of negative control wells (typically in the range 0.02−0.1).

The corrected score, or KS-score, measures the average KMG distance between a test compound and all negative control wells on the plate, with a correction for the "background" distance between control wells.

## RESULTS

**Choice of Cell Parameter.** A screening campaign was conducted for the 13 K test compounds, plus 4 Taxol positive controls, 4 ArQ positive controls, and 8 DMSO negative controls per plate, over a period of 11 days resulting in a total of ~35 K wells. The choice of which parameter from Table 1 to use in the comparison in eq 1 was made by a process of elimination and by set-wise comparison. A "good" choice of parameter in this case is one where (a) the average KMG distance between sets of control compounds (Taxol−Taxol, DMSO−DMSO, etc.) is close to zero and (b) the average KMG distance between known distinct compounds and the DMSO control wells (Taxol−DMSO, ArQ−DMSO) is significantly greater than zero.

Cursory inspection of Table 1 shows that a large number of parameters can be immediately eliminated because they are not expected to correlate with *any* distance measured with the Kolmogorov metric. These parameters include X, Y, X_size, Y_size, width, and height. More parameters were eliminated because of trivial transformations (e.g. Area**1/2 is simply the square root of area). The remaining parameters (Area, Perimeter, Wiggle, Integ-GS, Aver-GS, IQ-Range, Variance, StdDev, and AbsDev) were considered as candidates to rationalize the screening data. Table 2 shows the summary data computed from the screening campaign. Each

**Table 2:** Set-to-Set Kolmogorov Distances for Different Cell Parameters[a]

| set | parameter | MD(DD) | SD(DD) | MD(PP) | SD(PP) | MD(TT) | SD(TT) | MD(PD) | SD(PD) | MD(DT) | SD(DT) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | ABSDEV | 0.117 | 0.049 | 0.137 | 0.056 | 0.156 | 0.060 | 0.904 | 0.029 | 0.628 | 0.069 |
| A | AREA | 0.113 | 0.043 | 0.132 | 0.049 | 0.149 | 0.058 | 0.716 | 0.057 | *0.197* | 0.061 |
| **A** | **AVER-GS** | **0.124** | **0.054** | **0.160** | **0.073** | **0.156** | **0.060** | **0.890** | **0.035** | **0.597** | **0.072** |
| *A* | *INTEG-GS* | *0.117* | *0.045* | *0.131* | *0.044* | *0.151* | *0.061* | *0.186* | *0.055* | *0.480* | *0.077* |
| A | IQ-RANGE | 0.106 | 0.047 | 0.133 | 0.056 | 0.151 | 0.059 | 0.908 | 0.028 | 0.593 | 0.072 |
| *A* | *PERIMETER* | *0.096* | *0.038* | *0.117* | *0.049* | *0.135* | *0.054* | *0.735* | *0.051* | *0.159* | *0.049* |
| A | STDDEV | 0.117 | 0.049 | 0.136 | 0.056 | 0.156 | 0.059 | 0.898 | 0.030 | 0.632 | 0.069 |
| A | VARIANCE | 0.117 | 0.049 | 0.136 | 0.056 | 0.156 | 0.059 | 0.898 | 0.030 | 0.632 | 0.069 |
| *A* | *WIGGLE* | *0.114* | *0.044* | *0.129* | *0.048* | *0.153* | *0.059* | *0.684* | *0.062* | *0.210* | *0.070* |
| B | ABSDEV | 0.105 | 0.044 | 0.149 | 0.062 | 0.143 | 0.048 | 0.879 | 0.033 | 0.562 | 0.061 |
| B | AREA | 0.100 | 0.039 | 0.135 | 0.055 | 0.149 | 0.064 | 0.607 | 0.062 | 0.173 | 0.061 |
| **B** | **AVER-GS** | **0.105** | **0.042** | **0.181** | **0.086** | **0.135** | **0.051** | **0.867** | **0.038** | **0.521** | **0.064** |
| *B* | *INTEG-GS* | *0.104* | *0.041* | *0.127* | *0.048* | *0.145* | *0.055* | *0.150* | *0.053* | *0.395* | *0.068* |
| B | IQ-RANGE | 0.101 | 0.044 | 0.148 | 0.067 | 0.126 | 0.044 | 0.879 | 0.033 | 0.507 | 0.062 |
| *B* | *PERIMETER* | *0.091* | *0.037* | *0.128* | *0.058* | *0.138* | *0.062* | *0.586* | *0.064* | *0.165* | *0.057* |
| B | STDDEV | 0.104 | 0.043 | 0.153 | 0.064 | 0.161 | 0.058 | 0.873 | 0.034 | 0.571 | 0.061 |
| B | VARIANCE | 0.104 | 0.043 | 0.153 | 0.064 | 0.161 | 0.058 | 0.873 | 0.034 | 0.571 | 0.061 |
| *B* | *WIGGLE* | *0.104* | *0.042* | *0.129* | *0.051* | *0.151* | *0.064* | *0.651* | *0.059* | *0.170* | *0.068* |
| C | ABSDEV | 0.168 | 0.076 | 0.172 | 0.077 | 0.197 | 0.073 | 0.851 | 0.050 | 0.631 | 0.086 |
| C | AREA | 0.162 | 0.074 | 0.148 | 0.061 | 0.184 | 0.066 | 0.641 | 0.083 | *0.209* | 0.077 |
| **C** | **AVER-GS** | **0.172** | **0.078** | **0.175** | **0.079** | **0.201** | **0.077** | **0.833** | **0.054** | **0.613** | **0.087** |
| *C* | *INTEG-GS* | *0.169* | *0.076* | *0.165* | *0.068* | *0.200* | *0.077* | *0.261* | *0.096* | *0.478* | *0.109* |
| C | IQ-RANGE | 0.156 | 0.074 | 0.164 | 0.076 | 0.191 | 0.075 | 0.858 | 0.048 | 0.604 | 0.086 |
| *C* | *PERIMETER* | *0.147* | *0.069* | *0.134* | *0.061* | *0.170* | *0.062* | *0.638* | *0.079* | *0.185* | *0.066* |
| C | STDDEV | 0.168 | 0.077 | 0.171 | 0.077 | 0.198 | 0.073 | 0.842 | 0.052 | 0.632 | 0.085 |
| C | VARIANCE | 0.168 | 0.077 | 0.171 | 0.077 | 0.198 | 0.073 | 0.842 | 0.052 | 0.632 | 0.085 |
| *C* | *WIGGLE* | *0.170* | *0.079* | *0.156* | *0.063* | *0.193* | *0.073* | *0.664* | *0.091* | *0.234* | *0.095* |

[a] Where MD(XY) = mean KMG distance between compounds X and Y. SD(XY) = standard deviation of KMG distance between compounds X and Y. D = DMSO, P = proprietary compound ArQ, T = Taxol.

row is a different cell parameter used in the set-to-set distance computation taken on several days (days A, B, and C). Each column is a summary statistic for each set-to-set distance comparison. The mean interset Kolmogorov distance (without background correction) is shown in the table. The standard deviation of these distances is also reported. The summary statistics for subsequent screening days (days D−K) are not shown, but all show the same trends as shown in Table 2. See eq 2 (the mean interset Kolmogorov distance)

$$\mathrm{MD(A,B,}P) = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N_b}\mathrm{KMG}(P)_{ij}}{N_A * N_B} \qquad (2)$$

where MD(A,B,*P*) = the mean Kolmogorov distance between each compound in set A and set B using cell parameter *P*, $N_A$ = the number of wells in set A, $N_B$ = the number of wells in set B, $\mathrm{KMG}(P)_{ij}$ = the Kolmogorov distance between well *i* of set A and well *j* of set B using cell parameter *P*.

It can be seen from Table 2 that some parameters (area, INTEG-GS, PERIMETER, and WIGGLE) can immediately be eliminated because they fail to detect the difference between positive control compounds and negative controls. These are the italicized cells in the table. When the Kolmogorov distance is computed with these parameters, a small distance is computed, even when the images are different upon visual inspection. Thus five parameters remain that reliably measure a significant difference between negative and positive control wells, *and* no significant distance between cells that contain the same compounds. These parameters are ABSDEV, AVER-GS, IQ-RANGE, STDEV, and VARIANCE. In the rest of this paper the AVER-GS (average gray scale) parameter was used because of its simplicity and ease of interpretation.

**Visual Depiction of Screening Results.** With a choice of cell parameter made, a KS-score was computed for each test compound (for both replicates) by comparing it to negative controls using eq 1. It is important to note that the score is computed *locally*. The KMG distance was measured from each test compound to the negative control compounds on the *same plate*. This reduces errors that arise from small variations that may occur between plates. From the computed score and defining 0.2 as the cutoff, 30 significant hits were found. Each replicate was checked to confirm the hit. Note that the mean DMSO−DMSO (negative control) well distance is approximately 0.133 with a standard deviation of 0.058 (from Table 2 using the AVER-GS parameter). Rearranging eq 1 yields

$$\left(\frac{\sum_{i=1}^{N_D}\mathrm{KSS}_P(W,D_i)}{N_D}\right) = S(W,D,P) + B_C = 0.2 + 0.133 = 0.333$$

Thus a score cutoff of 0.2 plus the computed background correction of 0.133 implies that a compound must show an average KMG distance of at least 0.333 to be considered significant (3.4 standard deviations above the mean background correction value). Assuming normal statistics, only 0.2% of values are expected to lie more than 3 standard deviations from the mean. These are the "significant" hits.

The simple computation of KMG distances (and hence the score) is sufficient to identify the hits but not enough to fully understand the data. Once a distance can be computed between wells (the KMG distance or the corrected score), it is possible to find a projection of those wells into a 2D space such that the distances measured in 2D reproduce the true distances. A simple and elegant stochastic projection method has recently been published[19] that is scalable to large numbers

of points and easy to implement. Briefly, the algorithm proceeds by generating a set of random N−D (in this case 2D) coordinates for the wells and then making a large number of local pairwise corrections to the 2D coordinates such that the distances measured in 2D approximate the true distances. Pairs of points are selected at random, and their coordinates are adjusted according to

$$\left.\begin{aligned} \boldsymbol{P}_{A,i+1} &= \boldsymbol{P}_{A,i} + \left(\frac{\gamma^*(R_{AB} - D_{AB})}{(D_{AB} - \epsilon)}\right)*\left(\frac{\boldsymbol{P}_{A,i} - \boldsymbol{P}_{B,j}}{2}\right) \\ \boldsymbol{P}_{B,i+1} &= \boldsymbol{P}_{B,i} - \left(\frac{\gamma^*(R_{AB} - D_{AB})}{(D_{AB} - \epsilon)}\right)*\left(\frac{\boldsymbol{P}_{A,i} - \boldsymbol{P}_{B,j}}{2}\right) \end{aligned}\right\} \quad (3)$$

or where $\boldsymbol{P}_{A,i}$ = point $A$ at the $i$th inner iteration of projection, $\boldsymbol{P}_{B,i}$ = point $B$ at the $i$th inner iteration of projection, $R_{AB}$ = the actual distance between points $A$ and $B$, $D_{AB}$ = the current 2D distance computed from the projection, $\gamma$ = the scaling factor that is gradually reduced linearly during the projection, starting value 1.0, ending value 0.0, $\epsilon$ = the small correction factor to avoid overflow errors at $D_{AB} = 0.0$.

The algorithm proceeds over a fixed number of *outer* iterations (where $\gamma$ is reduced linearly from 1.0 to 0.0) and *inner* iterations (where eq 3 is applied for random pairs of points and a fixed value of $\gamma$). Full details of the algorithm are given in ref 19. All projections reported used 100 outer iterations and 5000 inner iterations.

Only two modifications to the protocol were made: the first allowed for the introduction of *fixed points*, whose coordinates did not change during the optimization. This is important for screening results because it might be useful to project test compounds into a space *already* defined by previous screening results (i.e. into a reference space of previously screened compounds). The projection equations are modified thus

$$\left.\begin{aligned} &\boldsymbol{P}_{A,i+1} = \boldsymbol{P}_{A,i} + \left(\frac{\rho^*\gamma^*(R_{AB} - D_{AB})}{(D_{AB} - \epsilon)}\right)*\left(\frac{\boldsymbol{P}_{A,i} - \boldsymbol{P}_{B,j}}{2}\right) \\ &\text{if point } \boldsymbol{P}_j \text{ is fixed, } \textit{otherwise} \\ &\boldsymbol{P}_{A,i+1} = \boldsymbol{P}_{A,i} + \left(\frac{\gamma^*(R_{AB} - D_{AB})}{(D_{AB} - \epsilon)}\right)*\left(\frac{\boldsymbol{P}_{A,i} - \boldsymbol{P}_{B,j}}{2}\right) \\ &\text{if both points can move, } \textit{or} \\ &\boldsymbol{P}_{B,i+1} = \boldsymbol{P}_{B,i} - \left(\frac{\gamma^*(R_{AB} - D_{AB})}{(D_{AB} - \epsilon)}\right)*\left(\frac{\boldsymbol{P}_{A,i} - \boldsymbol{P}_{B,j}}{2}\right) \end{aligned}\right\} \quad (4)$$

where $\rho$ = the asymmetry factor $1 <= \rho < 2.0$. Since only one point is moving it is permitted to move more than when the other point is moving also.

For the projections reported here, a value of 1.25 was used for the asymmetry factor, where appropriate.

The second modification introduced an estimate of the projection errors at the end of the projection. The errors were estimated for each well by selecting up to 100 random other wells and computing the mean of the difference between the real and projected distances.

**Interpretation of Screening Results.** 2D projections of the first 3 days (days A, B, and C) of screening are shown in Figure 3 using the KMG distance computed using the AVER-GS parameter. The choice of projecting all wells into 2D grouped by day is arbitrary. However, projecting each plate of results would mean more than 150 projections. Conversely, projecting the entire data set risks too many projection errors that would confuse the results. Taking results day-by-day is a compromise. The positive controls are Taxol (blue) and ArQ (red). The negative control is

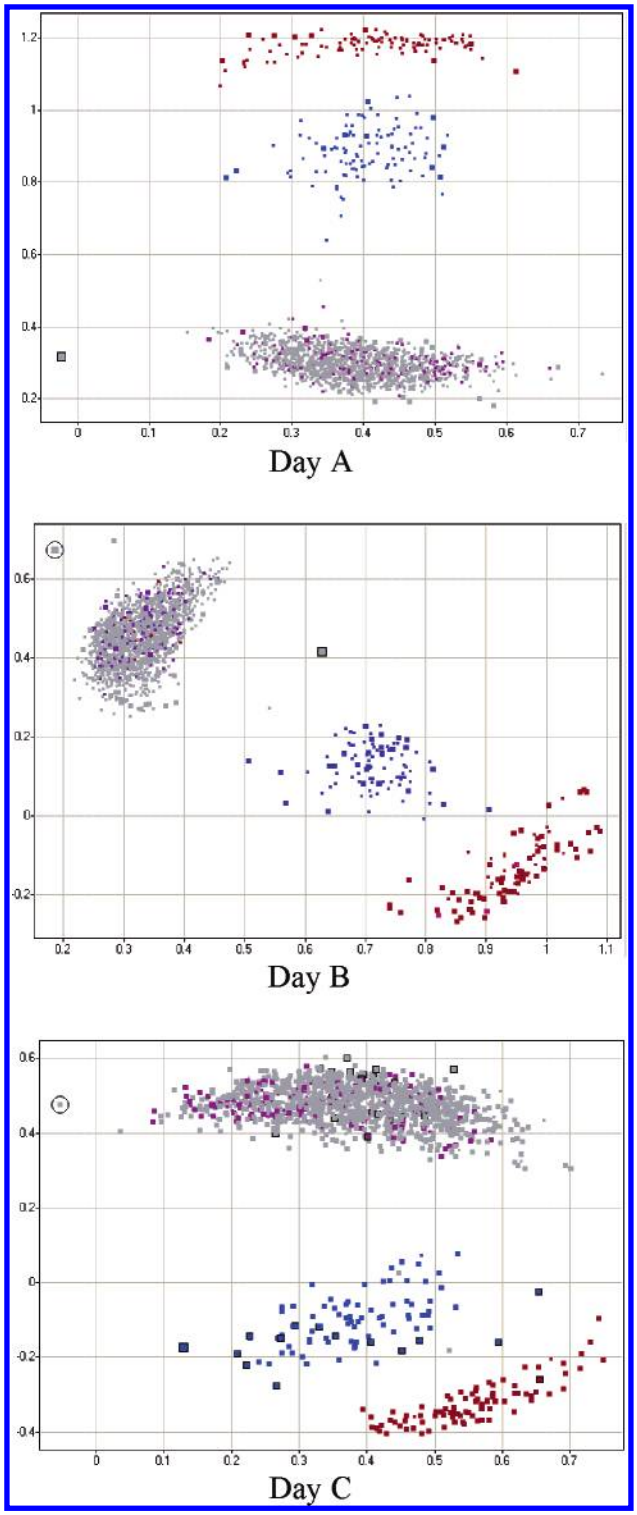


**Figure 3.** Projection of raw KMG distances on screening days A, B, and C 2D projections of wells. Red: proprietary compound wells; Blue: Taxol wells; Purple: DSMO wells; Grey: test wells. The size of the squares varies in proportion with the mean projection error.

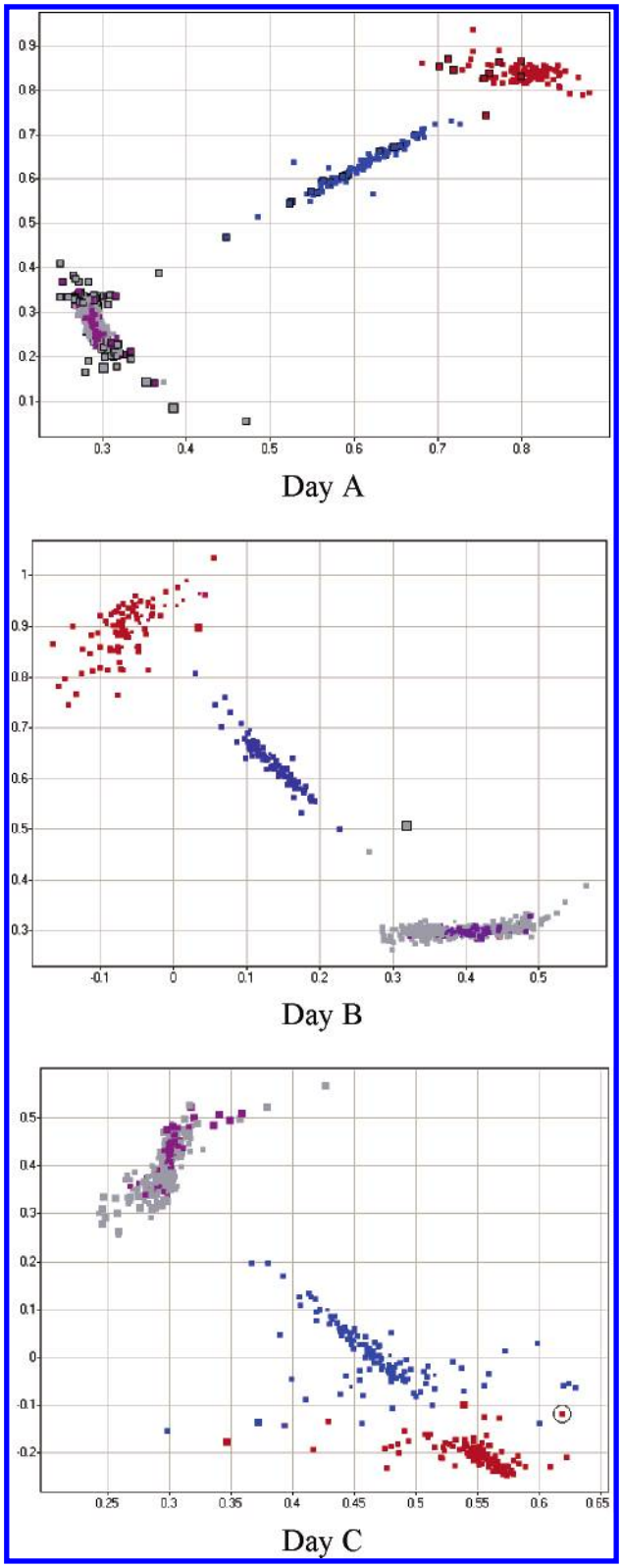


**Figure 4.** Projection of corrected KMG scores (eq 1) on screening days A, B, and C.

DMSO (purple). Screening compounds are shown in gray. The size of the points is proportional to the mean error of projection. Figure 4 shows the same data but with the background correction applied. Note, as expected, the cloud of DMSO points (purple) projected with the corrected score of eq 1 is much tighter than the raw distance projections. There is clear separation between the two positive controls

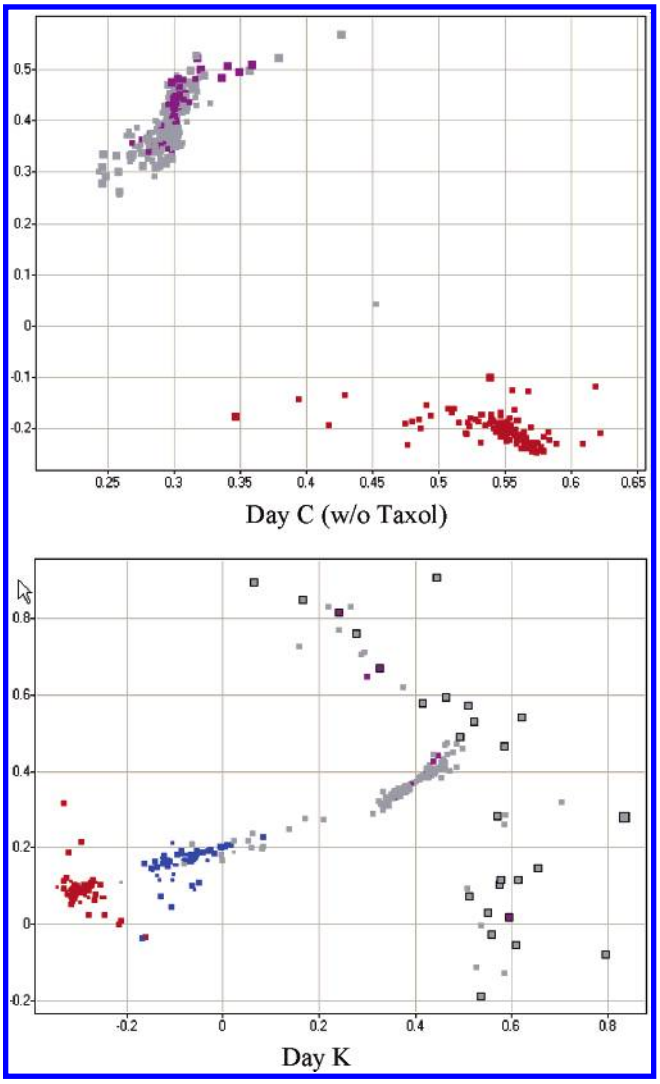


**Figure 5.** Depiction of hits.

(Taxol and ArQ) and the negative control (DMSO). Reassuringly the points (wells) with worse projection errors (i.e. the largest depicted points) generally lie at the edges of the clouds.

When comparing the projections from day-to-day, it is important to remember that all 2D plots have a rotational symmetry through any axis orthogonal to the plane of projection. The projected interwell 2D distances are invariant with respect to this rotation.

Each day of screening consisted of approximately 15 plates. On day C, some plates had problems with the contrast used in the imaging. This can clearly been seen from the projection in the form of more scattered points for the positive controls. This was the only projection that showed this pattern. This shows that projections are useful in quality assurance for the assays.

Figure 5 shows projections from day C and day K where genuine hits were found. The hits in day K were from a series that were expected to hit (by structural similarity to known hits), and these can clearly be seen in the projection, as the collection of gray points close to the Taxol (blue) cloud in Figure 5 (day K). The test compounds with high projection errors (i.e. large points) were ignored.

The raw images of some of wells are shown in Figure 2. Shown are representative images of a Taxol, ArQ, and

DMSO well from day C. Also shown is a hit, which was projected to be close to Taxol in Figure 5 (day C). This is the gray point in the center of the projection. Taxol was omitted from this projection for clarity. The DMSO image shows weak nuclear staining, indicating healthy cells. The images of ArQ and Taxol are distinctly different and show the cells have undergone a nuclear structural DNA change but in different ways. The Taxol staining pattern shows about 50% of the cells have undergone mitosis and display a punctate nuclear image (with tight, bright nuclear images). The ArQ staining pattern is different, with fewer very bright nuclear images. However, the ArQ image is different from DMSO, with the nuclei being contracted. The detected hit shows a pattern more similar to Taxol than to ArQ or DMSO. Both the mean Kolmogorov distance of this test compound to Taxol (~0.13) and the proximity of the compound to the Taxol cloud in the day K projection confirms the observation from the raw images.

## CONCLUSIONS

This paper has described a method for summarizing, rationalizing, and visualizing high-content screening data. The Kolmogorov statistic was used to compare populations of cells between wells, with a background correction applied. The choice of which parameter (computed from the gathering instrument) was made empirically using experimental data. A simple parameter was found (AVER-GS) that could place like compounds together (as measured by the Kolmogorov distance) and unlike compounds more distant. A background correction was made using the negative controls to derive a KS-score, and a threshold was defined to automatically find "hits" in the screening. A "hit" was defined as any test compound with a mean corrected Kolmogorov distance to all the controls of a plate greater than 0.2.

A fast and robust projection technique was used to depict the screening results from each day. Using these projections it was clearly seen that the screening score was able to differentiate the positive and negative controls. It was also seen that the projections gave good hints when the experimental data were noisy.

Using the KS score to rank test wells and the day-to-day projection to gauge overall assay quality, 30 compounds were selected for retesting over a larger concentration range (20 $\mu$M to 40 nM). Of these compounds 24 reconfirmed as positive in the DAPI assay. It is also noteworthy that in addition to screening the library using the staining assay a cell proliferation assay[5] was also performed. Of the 6 false-positive compounds, none were identified in the MTS screen and were suspected of being plate-handling errors prior to retesting in the DAPI assay. The overlap between the MTS and DAPI assay using the KS score metric coupled with the 2D projection demonstrates the robustness of the assay quality data processing tools described here.

## REFERENCES AND NOTES

(1) Giuliano, K. A.; Haskins, J. R.; Taylor, D. L. Advances in high content screening for drug discovery. *Assay Drug. Dev. Technol.* **2003**, *1*(4), 565−77.
(2) Mitchison, T. J. Small-molecule screening and profiling by using automated microscopy. *Chembiochem.* **2005**, *6* (1), 33−9.
(3) Abraham, V. C.; Taylor, D. L.; Haskins, J. R. High content screening applied to large scale cell biology. *Trends Biotechnol.* **2004**, *22* (1), 15−22.
(4) Vogt, A.; Cooley, K. A.; Brisson, M.; Tarpley, M. G.; Wipf, P.; Lazo, J. S. Cell-active dual specificity phosphatase inhibitors identified by high-content screening. *Chem. Biol.* **2003**, *10* (8), 733−42.
(5) Wilson, C. J.; Si, Y.; Thompson, C. M.; Smellie, A.; Ashwell, M.; Liu, J. F.; Ye, P.; Yohannes, D.; Ng, S. C. Identification of a Small Molecule that Induces Mitotic Arrest Using a Simplified High Content Screening Assay and Data Analysis Method. *J. BioMol. Screening* Submitted for publication.
(6) Yarrow, J. C.; Perlman, Z. E.; Westwood, N. J.; Mitchison, T. J. A high-throughput cell migration assay using scratch wound healing, a comparison of image-based readout methods. *BMC Biotechnol.* **2004**, *4* (1), 21.
(7) Yarrow, J. C.; Totsukawa, G.; Charras, G. T.; Mitchison, T. J. Screening for cell migration inhibitors via automated microscopy reveals a Rho-kinase inhibitor. *Chem. Biol.* **2005**, *12* (3), 385−95.
(8) Barak, L. S.; Ferguson, S. S.; Zhang, J.; Caron, M. G. A beta-arrestin/green fluorescent protein biosensor for detecting G protein-coupled receptor activation. *J. Biol. Chem.* **1997**, *272* (44), 27497−500.
(9) Wood, K. W.; Cornwell, W. D.; Jackson, J. R. Past and future of the mitotic spindle as an oncology target. *Curr. Opin. Pharmacol.* **2001**, *1* (4), 370−7.
(10) Mitchison, T. J. Small-molecule screening and profiling by using automated microscopy. *Chembiochem.* **2005**, *6* (1), 33−9.
(11) Perlman, Z. E.; Slack, M. D.; Feng, Y.; Mitchison, T. J.; Wu, L. F.; Altschuler, S. J. Multidimensional drug profiling by automated microscopy. *Science* **2004**, *306* (5699), 1194−8.
(12) For an example, see: Dudoit, S.; Fridlyand, J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.* **2002**, Jun 25, 3 (7).
(13) CLOGP was computed using ACD Labs Software, version 2.0041 available from www.acdlabs.com.
(14) LogS was computed with Pipeline Pilot, version 4.1 available from Scitegic at www.scitegic.com.
(15) Baldino, C. M.; Caserta, J.; Goetzinger, W.; Harris, M.; Hartsough, D.; Yohannes D.; Yu, L.; Kyranos, J. N. High-Throughput Medicinal Chemistry for Efficient Drug Discovery. *Curr. Drug Discovery* **2004**, *15*, 15−19.
(16) Kyranos, J. N.; Cai, H.; Zhang, B.; Goetzinger, W. K. High-throughput techniques for compound characterization and purification. *Curr. Opin. Drug Discovery Dev.* **2001**, *4* (6), 719−28.
(17) http://www.beckmancoulter.com
(18) Chakravart, Laha, Roy, *Handbook of Methods of Applied Statistics*; John Wiley: 1967; Vol. I, pp 392−394.
(19) Agrafiotis, D. K. Stochastic proximity embedding. *J. Comput. Chem.* **2003**, *24*, 1215−1221.

CI050404G