

Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds

Yan Feng^{*}, Timothy J. Mitchison[†], Andreas Bender[§], Daniel W. Young^{||} and John A. Tallarico^{*†}

Abstract | Multi-parameter phenotypic profiling of small molecules provides important insights into their mechanisms of action, as well as a systems level understanding of biological pathways and their responses to small molecule treatments. It therefore deserves more attention at an early step in the drug discovery pipeline. Here, we summarize the technologies that are currently in use for phenotypic profiling — including mRNA-, protein- and imaging-based multi-parameter profiling — in the drug discovery context. We think that an earlier integration of phenotypic profiling technologies, combined with effective experimental and *in silico* target identification approaches, can improve success rates of lead selection and optimization in the drug discovery process.

^{*}*Developmental and Molecular Pathways, Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.*

[†]*Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA.*

[§]*Medicinal Chemistry Division, Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands.*

^{||}*Wolf, Greenfield and Sacks, 600 Atlantic Avenue, Boston, Massachusetts 02210, USA.*

^{*}*Global Discovery Chemistry, Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.*

Correspondence to Y.F. or T.J.M.

e-mails:

yan.feng@novartis.com;

[timothy_mitchison@hms.](mailto:timothy_mitchison@hms.harvard.edu)

harvard.edu

doi:10.1038/nrd2876

Cell-based assays are increasingly being used in drug discovery to monitor drug responses of signalling pathways because they often more accurately reflect the complexity of the entire living organism than target-centric biochemical assays¹. However, directly applying cell-based assays to primary high-throughput screening is not without its own challenges. For example, identification of a hit compound often does not suggest a single, particular molecular target, preventing further optimization of hits by medicinal chemistry². This has led to cell-based screens being termed ‘black-box’ screens. In addition, cell-based screens tend to have higher hit rates than biochemical screens, presumably reflecting the existence of many potential targets in the intended biological pathway, or unexpected crosstalk from interconnected pathways and mild toxicity that give an activity read-out³. Even in a typical target-driven drug discovery scenario, potent compounds that are identified using biochemical assays can fail to generate the desired effect on cells or have undesirable side effects. Thus, there is an unmet need to triage hit compounds from cell-based screens for further development, and to rapidly evaluate cellular efficacy and pinpoint side effects of compounds from target-driven screens.

In the context of the ongoing paradigm shift towards pathway-driven drug discovery, it becomes increasingly important to accomplish these steps earlier in the drug discovery pipeline and with a greater throughput. Here, we propose multi-parameter cellular phenotypic

profiling as a promising solution. In this approach, broad and quantitative molecular and physiological measurements of cellular responses to compound treatment are used to provide information on compound activity and target mechanisms. We reason that this approach deserves more attention as an early step in the drug discovery pipeline.

The cellular phenotype is complex, as it is the summation of the activity state of many pathways. To reliably describe the phenotype, multiplexed measurements of those pathways are needed that can distinguish between phenotypes and at the level of throughput which is required early in the pipeline (typically hundreds to thousands of compounds per screen). There is often an inverse relationship between the level to which an assay is multiplexed and the throughput of the assay: whereas a high-throughput screen (HTS) with a single read-out can screen over a million compounds per week, highly multiplexed assays with tens of thousands of read-out parameters are often many magnitudes slower and are more costly (FIG. 1). In choosing appropriate profiling technologies, it is important to strike a balance between these two characteristics and to consider how the resulting data will be interpreted and used for compound stratification. Triage of HTS hits at an early stage, for example, requires high throughput, whereas studies to determine the mechanism of action and toxicity of the compound — for which there is a premium on mechanistic information — requires a greater range of read-outs (BOX 1).

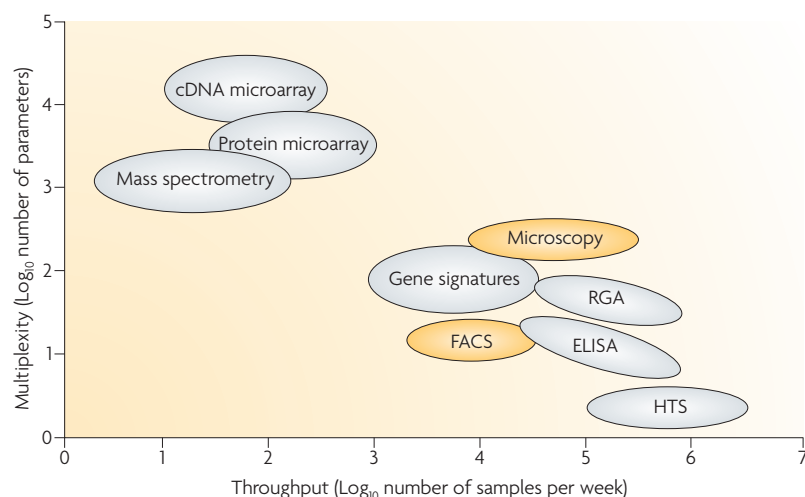


Figure 1 | Multiplexity and throughput of phenotypic profiling technologies.

The ability to multiplex and the throughput of the assay are often inversely correlated. Highly multiplexed assays with tens of thousands of read-out parameters, such as complementary DNA (cDNA) and protein microarrays, as well as protein mass spectrometry, are often many orders of magnitude slower than other less complex assays, such as enzyme-linked immunosorbent assays (ELISAs) or reporter gene assays (RGAs). ELISAs and RGAs, in single output format, are widely used as high-throughput screening (HTS) methods. Multiplexed ELISAs and RGAs, which are essentially many HTS assays running in parallel at lower throughput, can be used as profiling tools. Gene signature, imaging-based flow cytometry and microscopy assays have the potential to be simultaneously high throughput as well as highly multiplexed. Imaging-based assays (shown in yellow) also provide single-cell-based information for analysis of a heterogeneous cell population. FACS, fluorescence-activated cell sorting.

Several medium-throughput profiling approaches, such as gene signature- and image-based screens, can in principle provide rich mechanistic information at reasonable speed and cost. These technologies are continuously improving, so it is conceivable that it will soon be possible to generate multi-parameter activity profiles of entire screening compound collections (currently in the order of ~1–3 million compounds for a large pharmaceutical company). However, a fundamental challenge remains in understanding how to most informatively mine data that emerge from such multi-parameter approaches. In this Review, we summarize the current state of multi-parameter profiling technologies and discuss some of the emerging statistical tools that are being used to extract vital information from these approaches.

Multiplexing

A process by which many single-output assays are carried out in parallel.

Gene signature

A small subset of gene products (mRNA or protein) that are consistently upregulated or downregulated in certain disease states or following treatment with known drugs. The gene signatures can be used to mine expression profiling data for compound or disease association, or applied directly in higher-throughput detection methods, such as the quantitative polymerase chain reaction.

Multi-parameter profiling technologies

Multi-parameter phenotypic profiling includes transcriptional, proteomic and cell imaging-based measurements (FIG. 2). Each of these methods (summarized in TABLE 1) can simultaneously measure changes in several cellular processes that are relevant to both efficacy and toxicity in a cellular-system setting. Taken together, these methods therefore hold considerable promise for expediting drug discovery by identifying cellular processes that are amenable to targeting by small-molecule drugs. They also provide information that is potentially relevant to assessing the toxicity and mechanism of action of a compound.

Transcription. Transcription profiling using a complementary DNA (cDNA) microarray — which provides a global measurement of all the mRNA levels in a sample — was developed more than a decade ago and is now a standard technique^{4,5}. Early attempts to determine the mechanism of action of a drug using microarray data were reasonably successful^{6,7}. Although highly multiplexed, the application of transcription profiling in even medium-throughput contexts suffers from at least two major drawbacks: high cost and poor comparability between experimental data sets^{8–10}.

Currently, the cost of cDNA microarrays for full-genome transcript profiling remains prohibitively high for even a few hundred samples, owing in part to the large number of cells that are required and the low speed of the assays. Analysis of cDNA microarray data revealed that a small number of gene signature profiles are sufficient to classify compounds according to their known mechanisms of action^{6,11}. Thus, in practice, the higher-throughput gene signature-based methods, such as the quantitative polymerase chain reaction (QPCR)¹², mass spectrometry¹³ and the recently developed bead-based LMF (ligation-mediated amplification with Luminex-barcoded microsphere and flow cytometry detection) technology^{14,15}, are more suitable for large-scale compound profiling. Measuring a small subset of genes (5–100) has been successful in identifying relevant hit compounds in medium-throughput screens^{12,15}, indicating the reliability of those lower-dimension read-outs. The gene signature-based technologies have substantially driven down the cost of transcription profiling — by two orders of magnitude in some cases. It remains to be seen whether profiling against a selected subset of genes — such as those relevant to toxicity — will have a similar predictive power for mechanism of action analyses when more diverse compound classes are used.

Another transcription-based approach, the reporter gene assay (RGA), has been widely used for cell-based primary screening to identify pathway modulators^{1,3,16}. Individual RGA read-outs do not constitute a phenotypic profile, but many such assays performed in parallel have the potential to report on the biological activity of a compound. Thousands of predicted transcription factor binding sites are known, allowing a wide range of targets to be screened using RGA technologies¹⁷. Hundreds of artificial promoter-reporter gene assays have been developed over many years, and many are commercially available. Given their robust signal and amenability to high-throughput screening formats, multiplexed RGAs can readily be used for high-throughput phenotypic profiling of compounds.

The main concern in using RGAs to report on phenotypic effects is the use of artificially engineered reporters that may not reflect the endogenous regulation. Most physiological gene promoters integrate inputs from many signalling circuits, whereas artificial promoters — which often consist of tandem multimers of a particular transcription factor binding site — typically do not include much of this complexity or, even worse, introduce complexity in an uncontrolled and physiologically irrelevant way. The result is a trade-off, which on one

Box 1 | Comparative views of multi-parameter cell-based assays from two application angles

Each technology has advantages and disadvantages, and it is therefore important to be able to compare and choose the right technology for a particular purpose. Generally, multi-parameter cell-based assays are applied either as screening tools or as profiling tools for target identification, mechanism of action elucidation or toxicity studies. The purpose of the application will determine how each technology is evaluated.

Primary screens

From a screening point of view, the goal is to rapidly identify compounds that have a desirable profile and to eliminate those that do not. Screening assays should have reasonable throughput, simple analysis matrices to identify active compounds and robust criteria to select against undesirable side effects. Assays that require a large amount of source material or a slow, expensive or complex data format, such as complementary DNA (cDNA) or protein microarrays, and affinity-based proteomics by mass spectrometry are unlikely to be the methods of choice for screening. However, assays that can easily be miniaturized and automated and have a good signal to noise ratio — such as multiplexed reporter gene assays, enzyme-linked immunosorbent assays, gene expression signature assays, flow cytometry- or microscopy-based assays — can readily be applied to medium-throughput secondary screens or even primary screens. Traditionally, high-throughput primary screens (>1 million compounds) were designed with a single read-out to facilitate the selection of active compounds. For this approach, data from multi-parameter assays have to be reduced to single or several crucial parameters for hit selection. Whether to select hits for each parameter and then pool the hits for further analysis, or to reduce the multidimensional data into a single parameter to select a hit, has to be determined case by case. For example, in a gene signature screen with both upregulated and downregulated signature genes, should the direction of regulation affect the score? Because of these complicated issues, most multi-parameter assays, when applied to screening, are currently treated as a single (combined) parameter read-out assay. Other parameters are only used to indicate toxicity or to divide subpopulations of cells into different cell cycle phases or cell types to extract a single determining parameter. High-dimensional data can be particularly challenging to reduce into a biologically meaningful single parameter. Factor analysis, with the statistical power to identify underlying attributes, can be particularly useful for data reduction.

Secondary screens

By contrast, secondary screens operate on a smaller scale (several thousands of compounds) and are more flexible in terms of analysis. Here, the goal of multi-parameter profiling assays is to understand the mechanism of action or the toxicity of the compounds. Selected parameters with a particular biological meaning, alone or in combination, are usually visualized in a dose–response format by human inspection before decisions are made. In this case, multi-parameter profiling has the potential to provide more information pertaining to the mechanism of action of a compound than a biochemical screen. The potential mechanisms are usually derived from clustering with compound ‘landmarks’ which have known mechanisms of action on cells. A sufficient amount of standard control compounds in each assay is crucial to understanding which types of mechanism can be revealed by which assay format. Predicting compound binding activity on the basis of their chemical structures is complementary to mechanism of action prediction on the basis of experimental data. The type of assay selected will affect the power to predict certain mechanisms of action. For example, transcription-based assays are better suited to predicting which compounds will be general transcription modulators, as well as signal modulators that cause transcription changes; by contrast, imaging-based assays are better suited to predicting mechanisms related to morphological changes, such as cytoskeleton and membrane organization. Gene expression-based assays are better suited to identifying specific targets or pathways than are morphology-based assays. In principle, genomic assays with unbiased genome coverage, such as cDNA microarray or large-scale gene signature profiling and affinity proteomics, are conceptually superior to more biased imaging-based assays in terms of mechanism of action prediction, but less effective in identifying toxicities.

But in practice, measuring too many parameters can result in an over-determined system, and extensive data reduction is a prerequisite for effective and stable clustering analysis in many cases. Data reduction techniques — such as the Connectivity Map, sparse matrix factorization and factor analysis — can be helpful in managing such high-dimensional data. Another important aspect is that compound activities on cells are usually concentration-dependent and nonspecific toxicity usually occurs at high compound concentrations, whereas mechanisms of action should be concentration-independent. It is important to compare compounds with one another at their effective concentrations.

In general, there is no single technology that can solve all the screening and target identification problems. A sensible combination of high-throughput, simple assays and low-throughput, high-dimensional assays, as well as proper data integration, are the keys to expediting drug discovery.

hand provides a clear read-out of one pathway, but on the other hand excludes crosstalk and feedback from other pathways, which may be crucial to understanding phenotypic effects. Methods to improve physiological relevance, such as using relevant primary cells, might help improve the specificity of multiplexed RGAs.

Proteomics. Protein level and protein modifications — including phosphorylation, methylation, acetylation, ubiquitylation, sumoylation and palmylation — are important hallmarks of cellular signalling activities¹⁸.

Their accurate detection is therefore able to provide a detailed assessment of the biological state of a cell. Proteomic analysis is becoming an increasingly important tool in target and biomarker discovery. Enzyme-linked immunosorbent assays (ELISAs) or immunoblotting-based protein assays, on plates¹⁹, glass slide surfaces²⁰ or barcoded beads²¹, are widely used to monitor protein-state changes in HTSs. However, for finer resolution (that is, a higher multiplexity of read-outs) protein microarrays printed on glass slides with either capturing antibodies or cell lysates confer advantages²². Most highly multiplexed

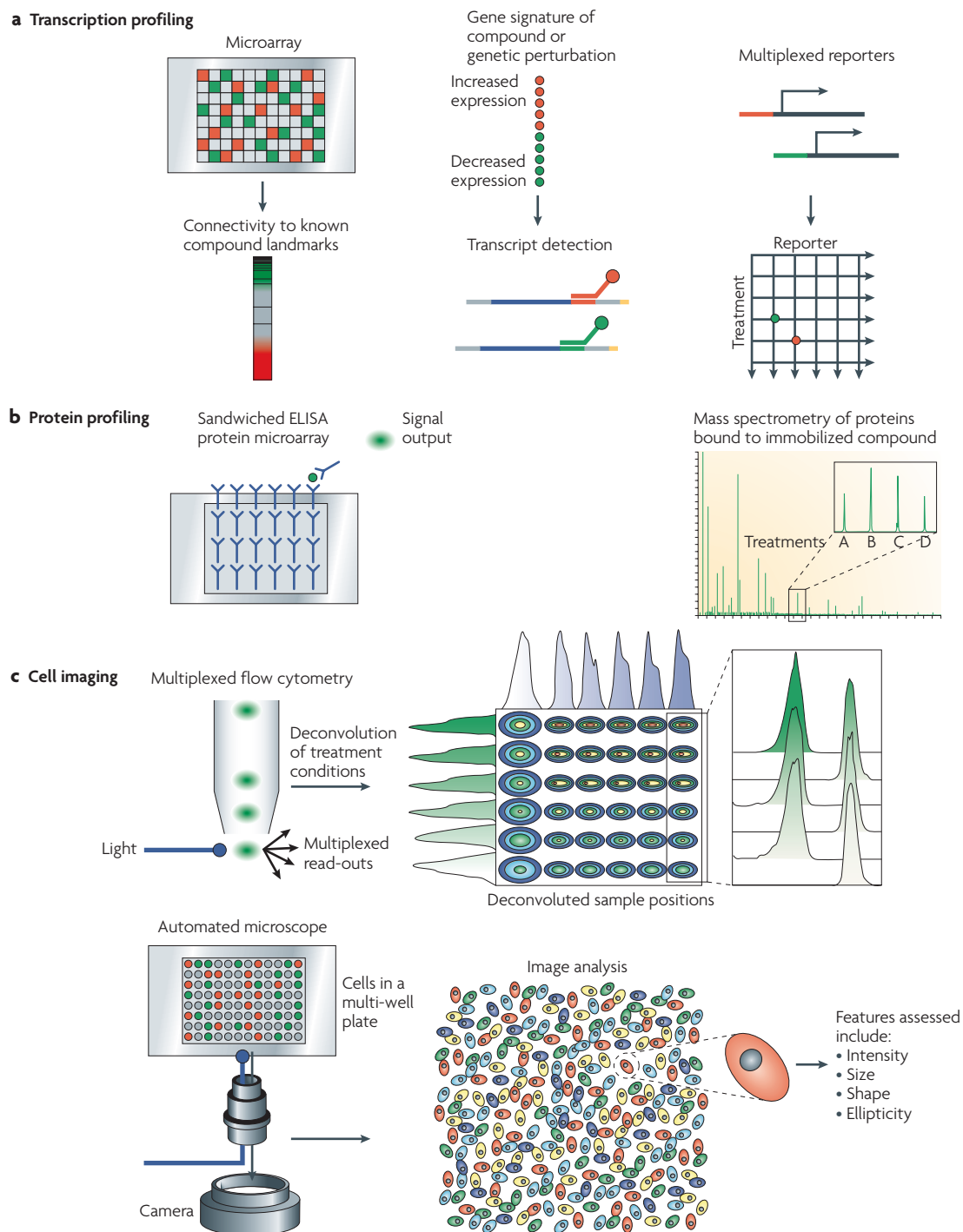


Figure 2 | Modalities of profiling technologies. **a** | Transcription profiles are monitored by whole-genome complementary DNA (cDNA) microarray assays, gene signature subsets and multiplexed reporter gene assays. An effective way to analyse complex data from whole-genome cDNA microarray assays is to measure 'connectivity' to existing 'landmarks' (green represents positive correlation and red represents negative correlation), whereas simpler data sets from small subsets of gene signature assays and multiplexed reporter gene assays (RGAs) can be directly subjected to clustering analysis as shown in the two-dimensional grid. **b** | Protein profiles are monitored by enzyme-linked immunosorbent assays (ELISAs), protein microarray or mass spectrometry. Sandwiched ELISAs and protein microarrays rely on antibodies to capture proteins of interest from the cell lysate. A second antibody that recognizes the same protein of interest, when used in detection, can usually increase the selectivity of the assay. Quantitative mass spectrometry, when used in compound affinity capture experiments, is particularly useful in identifying small changes in complex protein binding profiles produced by compound competition in cell lysates. **c** | Multiplexed flow cytometry can generate single-cell profiles on multiple proteins. One way to increase throughput for flow cytometry is to mix samples at the treatment level and separate the data after the read-out has been obtained, using position landmarks. High-content profiles can be generated by automated microscopy and image analysis.

Table 1 | Comparison of profiling technologies

Profiling technology	Overview	Advantages	Disadvantages
Complementary DNA microarray	Global measurement of all mRNA levels using microchip format	<ul style="list-style-type: none"> Highly multiplexed Established technology 	<ul style="list-style-type: none"> Small response window Large inter-experiment variation in signals Low throughput High cost
Gene signature	Selective analysis of subsets of signature mRNA levels by multiplexed QPCR or multiplexed ligation-mediated amplification and barcoded bead-based detection	<ul style="list-style-type: none"> High throughput Improved S–N ratio Lower cost than a complementary DNA microarray 	<ul style="list-style-type: none"> The predefined gene signature may not be closely associated with the relevant biology
Reporter gene	Expression of a reporter enzyme (for example, luciferase and β -lactamase), is driven by a multimer of transcription factor binding sites	<ul style="list-style-type: none"> Easy adaptation for high throughput Generally excellent S–N ratio Hundreds available for multiplexed profiling 	<ul style="list-style-type: none"> Artificially engineered reporters may not reflect the endogenous regulation
Antibody-based protein detection	Immunoblot, ELISA, antibody or lysate microarrays	<ul style="list-style-type: none"> Robust Sensitive detection Can be multiplexed 	<ul style="list-style-type: none"> Poor availability of validated reagents
Activity-based protein profiling by mass spectrometry	Global measurement of the binding of proteins in cell lysates to immobilized active compound	<ul style="list-style-type: none"> Detection of multiple binding targets is important for mechanism of action and toxicity inferences. 	<ul style="list-style-type: none"> Limited coverage Low throughput Requires many cells High instrumentation cost
Flow cytometry	Single-cell measurement of protein, DNA and ion content	<ul style="list-style-type: none"> Allows single-cell response to be monitored High S–N ratio Can be multiplexed 	<ul style="list-style-type: none"> Requires many cells Low throughput, but this can be overcome by an auto-sampler
Automated microscopy	Automated cell image acquisition and analysis	<ul style="list-style-type: none"> High spatial resolution High throughput 	<ul style="list-style-type: none"> Generates a large cellular descriptor data set, which creates challenges in finding biological relevance

ELISA, enzyme-linked immunosorbent assay; S–N, signal to noise; QPCR, quantitative polymerase chain reaction.

chip-based protein measurements depend on the availability of selective antibodies that detect the protein modification of interest. However, the development of antibodies to reliably measure protein modifications is still not adequately advanced²³ and so improved methods — for example, the use of directed evolution technologies such as phage displays — to make new detection reagents are needed. Specificity, which is often poorly validated, is another key area of uncertainty for many antibodies^{24,25}. Specificity can be improved by two-antibody sandwich methods or the use of certain fractionation methods in immunodetection, such as in traditional Western blots in which markers that correlate with molecular weight, to provide an additional specificity criterion. Although systems are now being developed for highly multiplexed protein microarrays and Western blots, progress in highly multiplexed protein detection has been limited by a lack of sufficient systems level knowledge and suitable detection reagents.

Protein mass spectrometry has recently enabled rapid improvements in fractionation methods and instrumentation that allows an increase in the analytical depth (the number of proteins identified in each sample) of protein mixtures. As a result, thousands of proteins can

be routinely detected in one experiment, which is ideal for profiling studies. Specific cellular protein–compound interactions can be surveyed by mass spectrometry using immobilized compound as ‘bait’ in activity-based protein profiling²⁶. The use of quantitative methods, such as iTRAQ (isobaric tag for relative and absolute quantitation) labelling²⁷, can also facilitate the identification of differential protein binding, especially when used in competition with soluble active compounds. For example, a study that used immobilized non-selective kinase inhibitors as a kinase-capturing tool revealed binding profiles of several well-known clinically used kinase inhibitors that closely correlated with known target activities. It was also able to identify the receptor tyrosine kinase DDR1 as a novel target of imatinib (Gleevec; Novartis)²⁸. The direct analysis of lysates of compound-treated cells is still hampered by under-sampling of the highly complex protein sample, which have vastly different dynamic ranges²⁹. Continuous improvement of multidimensional liquid chromatography³⁰ and adaptation of a targeted approach³¹ — by tuning the mass spectrometer to detect only peptides with predefined mass value, or ‘mass fingerprints’ — could enhance the sensitivity and reproducibility of proteomic profiling, leading to their routine use in

iTRAQ

(Isobaric tag for relative and absolute quantitation). A technique that uses isotope-coded covalent tags to quantify protein from different sources in one single mass spectrometry experiment.

assaying drug activities. Current limitations of proteomics include the low throughput for protein detection in complex mixtures of cell lysates by mass spectrometry and the upfront instrumentation set-up cost. We expect that the current interest in this area will lead to improvements in cost effectiveness in the near future.

Cell imaging. Although technologies to measure transcript and protein levels can be highly multiplexed, a limitation is that they measure an average across a population of treated cells, which decreases their sensitivity to phenotypic changes that affect subpopulations of cells. This is not only a problem in tissues and mixed cell cultures, but also in clonal cell lines, in which subpopulations frequently occur; furthermore, subpopulations might even be induced by drug treatment³².

Screening modalities that provide single-cell resolution, such as flow cytometry and microscopy, can overcome this drawback. Both automated microscopy and high-throughput flow cytometry make quantitative measurements at the single-cell level using multiple fluorescent channels. In general, flow cytometry tends to provide data on more cells because of its speed, and is highly sensitive and easier to multiplex (up to 14 channels) because of its simpler optical set-up. By contrast, microscopy-based read-outs tend to provide more information on each cell because of a superior spatial resolution. However, these differences are beginning to merge as flow cytometers with imaging capability are developed³³ and microscopy read-outs increase in speed. Certain practical differences will remain: for example it is easier to analyse non-adherent cells by flow cytometry and adherent cells by microscopy. In addition, microscopy-based assays are generally more suitable for high-throughput screening of multiple samples than cytometry-based assays because fixed, stained cells on plates can be stored for as long as needed before analysis.

In principle, flow cytometry can work in a high-throughput format. However, its ability to handle multiple small-volume samples is limited in most commercial instruments because the auto-sampler is slow (2–3 samples per minute). Technology improvements — such as the continuous-flow auto sampler, in which the entire multi-well plate of samples is treated as a single event separated by sampling time and air bubbles in between data points³⁴ — or mixing multiple samples that are fluorescently labelled according to their well positions before each analysis, have reduced the dead volume and number of cells required for each sample. This has enabled high-throughput flow cytometry experiments for profiling purposes³⁵ (FIG. 2).

Phosphorylation is a hallmark of intracellular signalling. The PhosphoFlow platform is a flow cytometry technology that simultaneously monitors several phosphorylation events in signal transduction pathways. In primary blood cells, specific cell surface markers can be used to simultaneously analyse specific T cell and B cell subpopulations³⁶. PhosphoFlow was able to detect correlations between parameters in each subpopulation of cells, whereas in lysate-based phosphoproteomics assays these important correlations were lost. Thus, in contrast

to well-average-based assays, PhosphoFlow can detect changes in small populations of cells³⁷. As with the gene expression signature³⁸, such cell-specific multi-phosphoprotein signatures could be used to predict therapeutic outcome by matching specific disease signatures to drug response signatures³⁹. Currently, the method can be used to screen several thousand compounds per day, a throughput sufficient for validation in a secondary screen and profiling in drug discovery^{37,40} (FIG. 2).

Compared with flow cytometry, fluorescence microscopy has a greater potential to be a high-throughput and highly multiplexed profiling tool. Only a small number of cells are needed to generate complex read-outs, which might be an important benefit in analysing precious human primary cells. Microscopy-based assays have the same advantage as flow cytometry-based assays in monitoring single-cell responses, and with suitable probes — such as green fluorescence protein — microscopy-based assays can also monitor live cellular responses over time⁴¹.

Recent advances in fluorescence microscopy technology have enabled the automated collection of images from multi-well plates, in three to four fluorescence channels, as a function of time. With appropriate analyses, this number of channels is sufficient to obtain large amounts of phenotypic information from a single well. This assay type is therefore known as a high-content assay (HCA)^{42,43}. Early HCAs only extracted one or a few parameters with the most obvious biological meaning: for example, nuclear translocation of a signalling molecule was used to report on the activity of a signalling circuit, and the number of viable cells in the wells was used as a measure of toxicity⁴⁴. Simple computer algorithms can generate many parameters with no apparent biological meaning. Currently, dozens of cell parameters are routinely measured by a common HCA, such as geometric cell dimensions, and intensity and texture features. A more complex analysis image can yield many more parameters⁴⁵. A typical HCA might generate gigabytes of data from the images that describe the amount and location of biomolecules in individual cells. High-content imaging can clearly provide data in quantities several orders of magnitudes higher than other profiling tools. However, in most cases, the precise information contained in each measurement, describing a complex biological observation that is often caused by several mechanisms, can be difficult to interpret. Certain parameters (known as descriptors), such as 'nuclear ellipticity' or 'variance in intensity', convey no biological meaning per se, in contrast to intracellular levels of mRNA transcripts or phosphorylated proteins. Yet, they might be an informative manifestation of the response of a complex underlying biological system to a perturbation, and changes in such parameters in response to a pharmacological agent provide a potentially useful way to conduct compound profiling. A major challenge for image-based phenotypic profiling has been to develop methods that can extract biologically relevant profiles of compounds from the huge data set available⁴⁶. Over the past 5 years, successful examples have emerged in which cytological profiles were used to categorize compounds with known activities and to identify novel mechanisms

High-content assay

An automated fluorescence microscopy-based assay that measures many parameters of cellular marker intensity and morphology pertaining to compound activity and toxicity.

Box 2 | Data normalization and analysis

Here we summarize recurrent themes in the analysis of data that are not restricted to phenotypic-screening data.

Normalization and standardization

First (and particularly if data are obtained from different sources), normalization or standardization of the data is important to allow for comparison of data — for example, from different laboratories, even if they are implementing the same protocols. The same is true for different plates in a screening deck, and for data recorded on different days or using different instruments. Normalization here refers to the linear scaling of all data into an interval between 0 and 1, whereas standardization projects all data points onto a distribution that is centred at 0, with a standard deviation of unity. In the case of normalization, the relationship between the data points is preserved (the value ratios remain constant), whereas in the case of standardization, points of different values are shifted relative to each other to achieve a normal distribution, potentially destroying some of the internal structure of the data.

Statistical tests

After the standardization or normalization of the data, statistical tests are used to determine whether events are sufficiently significant to support (or reject) a given hypothesis. If data closely follow a normal distribution or a log normal distribution, the Student's T Test or Z scores are typically used. If distributions are not normal, then non-parametric statistics such as the Kolmogorov–Smirnov test or rank-based tests, are often used.

Although in many cases the numerical values of the data points are used for further analysis, an alternative method is to use rank-based statistics — also referred to as distribution-free statistics — which recognize only the order of data points (instead of their numerical values). The use of rank-based statistics has been particularly useful for analysing transcription profiling data generated by cDNA microarrays. Here, the major problem of comparing microarray data from different experiments and sources has been that the large variation between experiments and between sample sources, not between compound treatments (which are the actual variable of interest), has become the driving force in determining how compounds are grouped (cluster definition).

Clustering

Another analysis method that is possible after data normalization or standardization assesses the clustering of the data points obtained, with the aim of revealing the underlying structure of the data set. This step addresses questions such as what types of clusters are contained in a set of phenotypic read-outs? Clustering analysis is thus used to explore the relationship between compound profiles; for this, hierarchical clustering is a commonly used method. This is an unsupervised clustering method to arrange similar data points (such as compounds) in 'tree-like' data structures, in which the proximity of connected data points represents the similarity of the phenotypic read-outs obtained. Control compounds of known mechanisms of action can be used to identify function clusters, with the assumption being that similar phenotypes are produced by identical mechanisms.

Visualization

The visualization of data points is particularly important because humans have tremendous capabilities to detect patterns in graphical objects. Whereas original data can be plotted, this is often difficult in practice because the data are multidimensional. Therefore, dimension reduction is commonly performed for visualization of data. The aim is to remove dimensions from the data, while retaining as much information of the original set as possible. Two broad classes of dimension reduction approaches exist: linear methods, such as principal component analysis (PCA) and factor analysis, and nonlinear methods, such as multidimensional scaling, which minimizes Kruskal's stress (that is, the difference in distance between two data points in original and scaled-down space). Whereas linear methods preserve an orthogonal coordinate system, nonlinear methods are often able to put data points into low-dimensional space into distances that better reflect their high-dimensional counterparts. In other fields, standard linear methods to analyse large, multidimensional, partially correlated data sets are mainly factor analysis and PCA. Factor analysis has perhaps been most commonly used to analyse market research data, but its features make it an attractive method to analyse almost any multivariate data set. The method is conceptually related to PCA but, rather than just accounting for the maximum total variance in a data set, factor analysis seeks to maximally account for the common underlying variance that is shared between variables. In certain contexts, in which the maximum total variance in a data set and the common underlying variance are at similar levels, the results of a PCA and factor analysis may converge.

Machine learning

As with visualization methods, machine learning (classification) methods for data classification also use linear and nonlinear dimension reduction methods. Whereas linear classification methods assume that features can be distinguished in an orthogonal coordinate system, nonlinear methods assume that there is a more complex relationship between input variables, such as the molecular structure of a ligand, and the output variable, such as the phenotypic read-out. This includes phenomena such as additivity of effects, as well as different output variable read-outs in different ranges of the input variable (for example, at different concentration intervals of a compound). Given the complexity of biological data, nonlinear analysis methods can often be better suited to the analysis of phenotypic screening data. Examples of this are artificial neural networks that try to emulate the human brain's way of learning, and support vector machines that project data points into a higher-dimensional space before linear separation of data points from different classes is performed. Overall, however, generating models for phenotypic read-outs is predominantly an empirical task, and it cannot be known beforehand which type of technique for generating activity models, for example, will give the best results.

Unsupervised clustering
A method used in machine learning to determine how data are organized without using predetermined training data.

Single-cell data

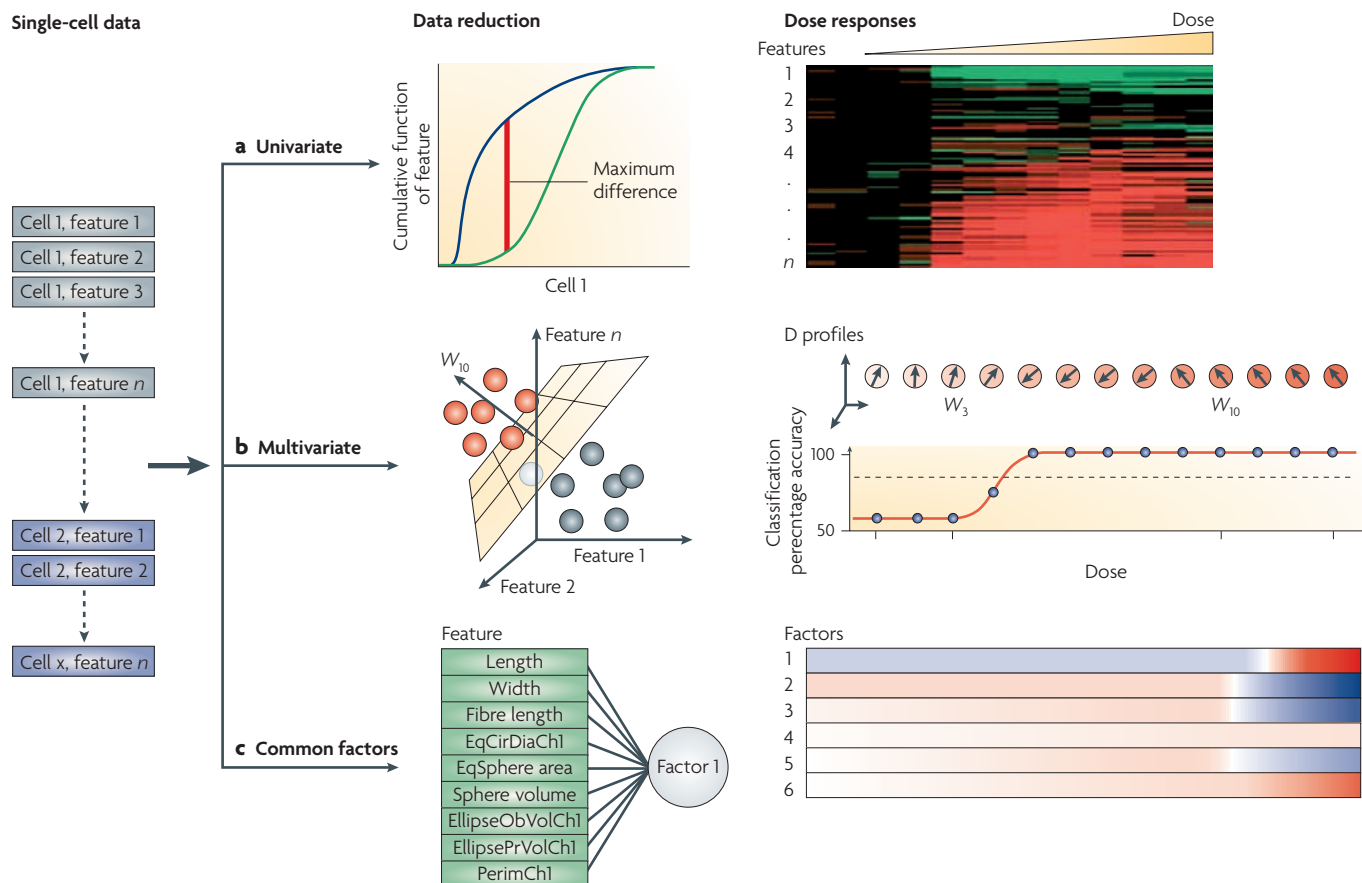


Figure 3 | Single-cell analysis. **a** | Univariate data analysis. Kolmogorov–Smirnov (KS) distance can be generated for each treatment condition by comparing treated and control cumulative functions. A ‘fingerprint’ of each compound’s activity can be compiled with KS scores of all features over a wide concentration range. **b** | Multivariate data analysis. Control (grey) and treated (red) cells can be separated by a support vector machine (SVM) classifier that is trained to recognize all available single-cell features. Classifier vector (D profile) denotes compound mechanism of action (W_3) and toxicity (W_{10}), and classification accuracy denotes the concentration–response relationship. **c** | Factor analysis. All features can be linearly combined to identify major underlying common factors (features contributing to factor 1 are shown). Major factors can be used to express both the mechanism of action (red represented upregulated and blue represented downregulated) and concentration responses. Part **a** is modified, with permission, from REF. 48 © (2004) American Association for the Advancement of Science. Part **b** is modified, with permission, from *Nature Methods* REF. 59 © (2007) Macmillan Publishers Ltd. All rights reserved.

pertaining to efficacy and toxicity^{47–49}. Live-cell microscopy remains an interesting area in which high-throughput automated data capture and image analysis are under development^{50,51}. It is well known that treatments can cause a rapid change in intracellular signalling. However, it remains to be seen precisely how live-cell imaging can be applied to compound profiling.

Data analysis

The goals of data analysis in phenotypic profiling include stratification of hit compounds, the identification of the mechanism of action of the compound at the pathway level and the elucidation of toxic mechanisms. Stratification of hits in multidimensional space can be a difficult task, in contrast to the usual rank ordering by efficacy or potency in a single-read-out assay. Our group recently developed a method for hit stratification that involves reducing high-dimensional multiplex image-based data into a series of one-dimensional data⁵². Phenotypic changes in any one or

more dimensions are projected onto a single dimension — for example, the level of cell cycle arrest or death pathway activation. Using this method, compounds that are bioactive with respect to one particular phenotype of interest can be identified. Subsequently, several phenotypic effects of each compound are analysed to give a multidimensional profile, in which each dimension corresponds to a specific phenotypic effect (such as cell cycle arrest), rather than a primary measurement (such as nuclear ellipticity), which is difficult to interpret. Target or mechanism of action predictions can be made by comparing the phenotypic profiles of new compounds with reference bioactive compounds, which have known targets or mechanisms of action and serve as landmarks in the multidimensional space⁵¹. Statistical tools for data normalization, clustering (that is, grouping compounds that cause similar phenotypic changes, and may or may not share chemical similarity) and visualization of multidimensional data are crucial for this type of analysis (BOX 2).

Box 3 | The challenges to effective phenotypic profiling

Here, we outline what we see as the most important current challenges in phenotypic profiling:

- Successfully combining information-rich 'omic data sets generated by current high-throughput technologies.
- Implementing profiling on relevant cell types, including but not limited to three-dimensional culture and cells differentiated from stem cells containing disease-relevant DNA sequence variants.
- Using new data-mining methods and finding ways to extract system level information from both drug-induced and genetic perturbations.
- Developing technologies to acquire and handle single-cell and time-dependent information.

Two distinct types of data can be generated by multi-parameter profiling technologies: well-based read-outs (which give the average response of a cell population that is subjected to one treatment) and single-cell-based read-outs (which provide measurements derived from each individual cell in the treated population). Transcription and proteomics profiling average entire wells rather than individual cells, and thus produce only well-based data. Fluorescence-activated cell sorting (FACS) and microscopy generate descriptors for every cell, of which there can be thousands, for each treatment. These can be averaged for a well-based read-out or analysed on a single-cell basis. Averaging will remove information, particularly as cell responses are heterogeneous. Single-cell analysis is therefore preferable in principle, but creates a burden in data management and analysis that challenges current hardware and software.

Analysis of well-based (population-averaged) data. Whereas data produced by multiplexed, single-parameter assays, such as multiplexed RGAs or ELISAs, are usually independently normalized and then clustered to identify groups of active compounds (BOX 1), multi-parameter well-based assays, such as cDNA microarray assays, require more attention during the data-handling stages. The application of a range of normalization, clustering, dimension reduction and machine learning techniques to multi-parameter cDNA microarray data has been well documented and reviewed⁴. A particular problem for this data is that the large variation between experiments and between sample sources has become the driving force in determining how compounds are grouped (cluster definition), presumably owing to the low signal to noise ratio of the data and its unusually high number of dimensions (in the order of tens of thousands of genes)⁵³. The use of rank-based statistics, such as in the 'Connectivity Map', has been particularly successful in improving the accuracy of prediction of compound mechanism of action and reducing the effect of experimental variations⁵⁴. In a series of recent publications, the Connectivity Map approach was able to correctly predict mechanisms of action of histone deacetylase inhibitors, oestrogen receptor modulators and antipsychotics, using data from separate experiments and various drug concentrations. This approach was used to predict that the unknown target of gedunin is heat shock protein 90 (REF. 55) and

to estimate the potential rapamycin sensitivity of a dexamethasone-resistant form of leukaemia³⁸. Remarkably, these predictions were both subsequently confirmed experimentally. In principle, it could be beneficial to apply rank-based statistics to other high-dimensional data sets (with thousands of parameters), such as those generated by activity-based protein profiling, and to apply statistical methods developed for multiplexed single parameter assays to lower-dimensional data (with only tens of parameters), such as gene signatures.

Analysis of single-cell-based data. In contrast to other profiling technologies, FACS and microscopy read-outs produce phenotypic profiles of compounds in single cells. This has the advantage that compounds which are active on only a subpopulation of cells can be observed. For example, T cell and B cell populations or cells that are in different cell cycle phases can respond differently to compound treatments^{35,56–58}. In the studies cited here, subpopulations were defined by well-known biological markers such as DNA content or cell surface markers. But in practice, subpopulations devoid of immediate 'biological meaning' could also be retrospectively defined by other parameters collected during the profiling experiment⁵⁷. The compound profile is defined by the combination of response parameters of the various subpopulations of cells.

In general, single-cell data dramatically complicate data analysis. A standard approach for reducing population data into treatment profile data has not yet been established. This means that each parameter is analysed either independently using a univariate data analysis approach or in combination using multivariate approaches.

In a pilot study to test whether a range of morphological parameters extracted from cell images contains information pertaining to drug mechanism of action, a non-parametric Kolmogorov–Smirnov score (KS score) of each parameter was generated by comparing cumulative functions of drug-treated cells with that of control cells⁴⁸. In this case, the KS scores represent a summarized treatment-based parameter extracted from thousands of responding cells. KS scores for all the features were used to generate a fingerprint-like signature profile of the compound (FIG. 3). The method was successful in clustering several classes of drugs, such as DNA-damaging agents, histone deacetylase inhibitors and microtubule poisons. Interestingly, selective omission of spatial parameters from the HCA data to mimic other intensity-based assays, such as flow cytometry (single cell) and cytoblot (whole-population average), failed to generate meaningful clusters in more than half of the compounds with a known mechanism of action, in contrast to the HCA. This shows that spatial features are essential for accurate classification.

The approaches discussed so far use single-cell data followed by univariate data reduction to generate a well-based measurement for each parameter. This tends to reduce the ability to classify subpopulations. A more complete analysis would consider all the data for a particular cell and include information from the correlation of parameters at the single-cell level, forming a truly multivariate analysis.

Kolmogorov–Smirnov score

The minimum difference between the empirical distribution function of the sample and the cumulative distribution function of the reference.

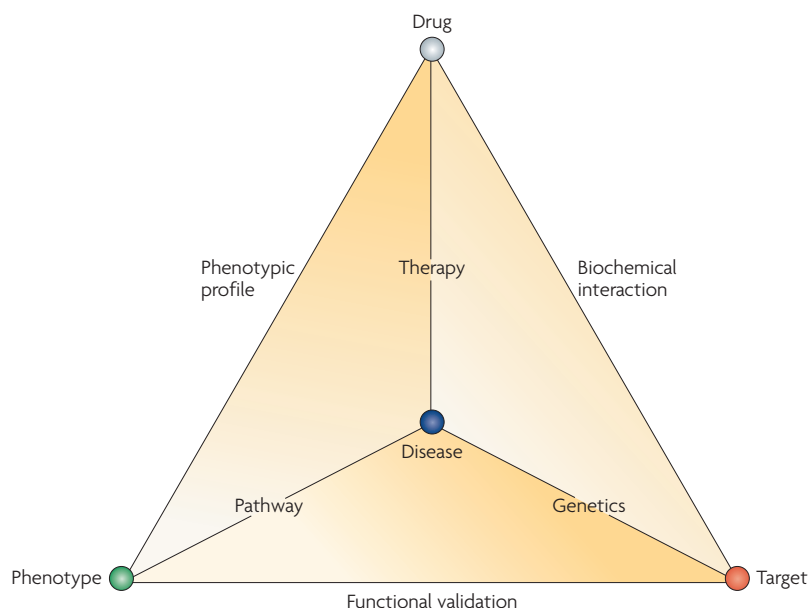


Figure 4 | Integration of information. Properly integrated genetic, biochemical, pathway and functional information is crucial for developing effective therapy for disease. Phenotypic profiling links drug and disease phenotype and so provides information pertaining to targets that can be modulated by drugs. The most important challenge is integrating phenotypic profiles with genetic mutations, chemical similarity and biochemical activity profiles in a way that improves decision making and enables rapid hypothesis and model testing. Systematic integration of phenotypic approaches into the existing approach that moves from disease target to drug is predicted to improve success rates of lead selection and optimization.

Support vector machine (SVM). A supervised learning algorithm, which is used for classification. An SVM constructs a separating hyperplane to maximize the difference between the treated and control cell data sets.

Factor analysis

A statistical method that explains the variability of observed variables in terms of reduced numbers of unobserved variables called common factors. The observed variables are modelled as linear combinations of the factors, plus 'error'. The factors often carry more interpretable meaning than the observed variables themselves.

Principal component analysis

(PCA). A data transformation method that is used to reduce multidimensional data sets to lower dimensions for analysis. PCA can reveal the internal structure of the data in a way that best explains the variance in the data.

One approach to overcome this challenge is to simply throw all the single-cell data into an n -dimensional analysis space, and then look for hyperplanes that separate them into statistically justified subpopulations (FIG. 3b). This approach is sensitive to heterogeneity in cell responses: indeed, quantification of heterogeneity is one of the main outputs of the analysis. Application of this approach to the same data compendium that was generated in the above-mentioned pilot study in which well-based KS statistics were used substantially increased the drug classification coverage and accuracy^{57,59}. In one of these studies, a support vector machine (SVM) classifier was trained to separate treated cells from control cells⁵⁹. The accuracy of cell classification in each treated well was used as an indicator of the level of response and the classifiers were used as indicators of distinct cellular responses to compound treatments (FIG. 3). This kind of analysis, in which all the data from single cells are explored, seems a promising approach to untangle the complexity of drug responses at the cellular level. It is also plausible that this kind of single-cell profiling will be applied to patient biopsies to develop a form of high-content histology that might provide a more informative diagnosis and patient stratification in diseases such as cancer.

In multi-parameter assays, many parameters reflect a similar underlying biology, and those read-outs are usually closely correlated with one another. Standard methods to analyse large, multidimensional, partially correlated data sets are factor analysis and principle component analysis (PCA)⁶⁰ (BOX 2). Such approaches are useful for

analysing phenotypic data because they allow a large data reduction but retain most of the information content and yield data-derived factors that are mechanistically interpretable by inspecting the parameters that contribute to each factor in many cases.

When factor analysis was recently applied to a data set of high-content compound screening, it rapidly reduced the data from 2×10^8 primary measurements to ~80,000 factor measurements⁵². A simple distant measurement that was formed on the basis of the factors was sufficient to identify hit compounds in the screen. Clustering of compounds according to a small set of independent factors, in combination with structural fingerprints of the compounds, successfully elucidated a structure–activity relationship. Therefore, the dramatic data reduction and simplicity of factor analysis is uniquely suited to both large-scale screening experiments and mechanistic profiling of active compounds.

Integration of information

When applied to collections of compounds, all of the phenotypic profiling approaches discussed generate meaningful clusters. A fundamental limitation is that these clusters typically do not, by themselves, reveal the biological mechanism of action of a compound, owing to the descriptive, empirical nature of the phenotypic read-out. To infer mechanisms, compounds with known mechanisms of action are typically used as activity landmarks. This comparative approach suffers from two drawbacks. First, the number of landmark compounds with a clear and specific mechanism of action in cells is limited. Second, many drugs have several targets and thus multiple mechanisms of action. How might we move away from the reliance on reference compounds, and towards a truly multi-parameter read-out of phenotype?

One approach is to integrate measurements of chemical similarity with phenotypic data. By training statistical models using compound structures that are associated with known targets^{61–63}, new relationships can be revealed between compounds and phenotypes. Another approach is to integrate knowledge on signalling pathways and disease into the analysis⁶⁴. For example, compounds that perturb different targets in the same pathway are expected to cause similar phenotypic changes. Interestingly, the statistical relationship between a high-content profiling read-out and the predicted protein target was found to be more significant than the relationship between a high-content profiling read-out and the chemical similarity of the compound to other compounds in the screen⁵², underscoring the target prediction power of phenotypic profiling. In future, by combining *in silico* target prediction steps with phenotypic profile-based genetic screening^{65,66}, phenotypic profiles could serve as a common link between the bioactivity space covered by the chemistry of a compound and the biological space modulated by its targets in the cell (FIG. 4). Another fascinating example of data integration was the use of a database of side effects of approved drugs together with measurements of chemical similarity to infer off-target drug interactions in humans⁶⁷. There remains great scope for progress in the area of data integration,

Structure–activity relationship

A correlation constructed between the features of chemical structures in a set of candidate compounds and parameters of biological activity, such as potency, selectivity and toxicity.

and it is clear that we are not yet using all the available knowledge to understand how compounds affect cells and thus stratify hits in the most effective way.

Conclusions and perspectives

The initial leads of many therapeutic drugs on the market today originated from some kind of phenotypic information, often whole-organ or whole-animal pharmacology or even whole-human data (for example, herbal medicines)⁶⁸. Rapid expansion of genomic information and high-throughput screening technology has increasingly pushed the drug discovery process into a rigid ‘target-driven’ approach that relies on poorly validated targets or simplified disease models for primary screening. The extent to which this shift is responsible for decreased productivity and soaring costs of drug development is debatable^{10,69}; our current understanding of how potential targets function at the system level is still poor, so perhaps it is inevitable that target-based approaches will often result in failure⁷⁰. Given the current lacklustre drug pipelines, a partial return to phenotypic approaches, empowered by modern screening technology, is worth considering.

We reason that cells — as the smallest living system — together with more complex model systems, such as tissues, should be more widely used to identify effective drugs. The challenge is how to efficiently use the multi-faceted phenotypic information we can gain from cell-based assays to increase the efficacy and reduce the toxicity of novel drugs. Combining these data-rich technologies should help to overcome the current challenge of predicting the biological effects of drugs before clinical trials. We foresee a continuous improvement of methodologies to enable the application of phenotypic profiling at an earlier stage and more widely in the drug discovery process (BOX 3). It will be important to implement these methods on relevant cell types and make use of developments in cell-based assays. These developments include three-dimensional cell cultures that more effectively mimic the tissue environment^{71,72} and the creation of important cell types by differentiating stem cells⁷³, including patient-derived induced pluripotent stem cells that recapitulate genetic drivers of disease⁷⁴. Another frontier in the field is new data-mining methods, as well as ways to extract system level information after both drug-induced and genetic perturbation, and integrate other sources of relevant information, including

measurements of chemical structure, signalling pathways and disease models, and effects of known medicines^{63,75–77}. We anticipate that several aspects of phenotypic profiling, such as single-cell information and possibly phenotypic changes over time, will be treated as important parts of the drug discovery puzzle in the future^{78,79}.

One particularly important question for the future of cell-based assays and phenotypic profiling is whether a biological target must be identified before a drug discovery programme can progress to the clinic. Historically, this was not required, and to some extent current efforts to develop kinase inhibitors and central nervous system receptor inhibitors with multiple targets represent a new approach to the classical target-free, phenotypic approach⁸⁰. We contend that if a phenotype of interest can be precisely defined, cell-based assays may be sufficient for preclinical development up to animal testing, and quantitative phenotypic profiles can be used to effectively predict structure–activity relationships. Although the cell-based structure–activity relationship obtained in this way may not be as precise as that obtained from biochemical assays, it is the structure–activity relationship from a therapeutic viewpoint that matters. Knowing a biological target will accelerate the hit-to-lead phase, for example, by allowing structure-based optimization of affinity, but this may not be a limiting factor in the overall development timeline. Identifying the targets of biologically active natural products often led to increased knowledge, but the initial natural product, or simple derivatives that could have been (and often were) made without knowledge of the target, remained the best lead. However, improvement of the compound often leads to identification of new drugs that act on the same target⁸¹, and technologies that allow rapid target inference have been under constant development⁸². There is nothing wrong with target identification per se — the question is whether it adds as much value as the popularity of the target-centric drug discovery paradigm would suggest.

To summarize, the wealth of current phenotypic read-outs that we have at our disposal, combined with powerful target identification, should provide insight into new processes by which to discover high-efficacy and low-toxicity therapeutics. Questioning the centrality of protein targets in drug development may be heterodox, but disease is an aberrant phenotype, not an aberrant pathway or target. Perhaps a phenotypic approach to identify new therapeutics is therefore needed.

1. Fishman, M. C. & Porter, J. A. Pharmaceuticals: a new grammar for drug discovery. *Nature* **437**, 491–493 (2005).
2. Hart, C. P. Finding the target after screening the phenotype. *Drug Discov. Today* **10**, 513–519 (2005).
3. Crisman, T. J. *et al.* Understanding false positives in reporter gene assays: *in silico* chemogenomics approaches to prioritize cell-based HTS data. *J. Chem. Inf. Model* **47**, 1319–1327 (2007).
4. Butte, A. The use and analysis of microarray data. *Nature Rev. Drug Discov.* **1**, 951–960 (2002).
5. Yang, Y. H. & Speed, T. Design issues for cDNA microarray experiments. *Nature Rev. Genet.* **3**, 579–588 (2002).
6. Gunther, E. C., Stone, D. J., Gerwien, R. W., Bento, P. & Heyes, M. P. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles *in vitro*. *Proc. Natl Acad. Sci. USA* **100**, 9608–9613 (2003).

- This study describes the development of statistical methods to classify drugs using expression profiling data and the discovery of a small set of expression biomarkers for classifying antipsychotic drugs.**
7. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
 8. Bugelski, P. J. Gene expression profiling for pharmaceutical toxicology screening. *Curr. Opin. Drug Discov. Devel.* **5**, 79–89 (2002).
 9. Butcher, R. A. & Schreiber, S. L. Using genome-wide transcriptional profiling to elucidate small-molecule mechanism. *Curr. Opin. Chem. Biol.* **9**, 25–30 (2005).

10. Stoughton, R. B. & Friend, S. H. How molecular profiling could revolutionize drug discovery. *Nature Rev. Drug Discov.* **4**, 345–350 (2005).
11. Gunther, E. C., Stone, D. J., Rothberg, J. M. & Gerwien, R. W. A quantitative genomic expression analysis platform for multiplexed *in vitro* prediction of drug action. *Pharmacogenomics J.* **5**, 126–134 (2005).
12. Bol, D. & Ebner, R. Gene expression profiling in the discovery, optimization and development of novel drugs: one universal screening platform. *Pharmacogenomics* **7**, 227–235 (2006).
13. Stegmaier, K. *et al.* Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nature Genet.* **36**, 257–263 (2004).
14. Peck, D. *et al.* A method for high-throughput gene expression signature analysis. *Genome Biol.* **7**, R61 (2006).

- This study describes a low-cost and scalable approach to gene expression signature analysis that combines ligation-mediated amplification with an optically addressed microsphere and a flow cytometric detection system.**
15. Stegmaier, K. *et al.* Signature-based small molecule screening identifies cytosine arabinoside as an EWS/FLI1 modulator in Ewing sarcoma. *PLoS Med.* **4**, e122 (2007).
- The work described here showed that gene expression signature profiles can be used in a primary screen to identify lead compounds for previously untractable targets.**
16. Bronstein, I., Fortin, J., Stanley, P. E., Stewart, G. S. & Kricka, L. J. Chemiluminescent and bioluminescent reporter gene assays. *Anal. Biochem.* **219**, 169–181 (1994).
 17. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
 18. Walsh, C. T. *Posttranslational Modification of Proteins: Expanding Nature's Inventory* 1–47 (Roberts and Company, Greenwood Village, 2005).
 19. Stockwell, B. R., Haggarty, S. J. & Schreiber, S. L. High-throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications. *Chem. Biol.* **6**, 71–83 (1999).
 20. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
 21. Earley, M. C. *et al.* Report from a workshop on multianalyte microsphere assays. *Cytometry* **50**, 239–242 (2002).
 22. Gembitsky, D. S., Lawlor, K., Jacovina, A., Yaneva, M. & Tempst, P. A prototype antibody microarray platform to monitor changes in protein tyrosine phosphorylation. *Mol. Cell Proteomics* **3**, 1102–1118 (2004).
 23. Zhu, H. & Snyder, M. Protein chip technology. *Curr. Opin. Chem. Biol.* **7**, 55–63 (2003).
 24. Haab, B. B., Dunham, M. J. & Brown, P. O. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* **2**, RESEARCH0004 (2001).
 25. Michaud, G. A. *et al.* Analyzing antibody specificity with whole proteome microarrays. *Nature Biotechnol.* **21**, 1509–1512 (2003).
 26. Cravatt, B. F., Wright, A. T. & Kozarich, J. W. Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. *Annu. Rev. Biochem.* **77**, 383–414 (2008).
 27. Chong, P. K., Gan, C. S., Pham, T. K. & Wright, P. C. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: implication of multiple injections. *J. Proteome Res.* **5**, 1232–1240 (2006).
 28. Bantscheff, M. *et al.* Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature Biotechnol.* **25**, 1035–1044 (2007).
- This paper describes a quantitative proteomics method to generate a binding profile for a kinase inhibitor.**
29. Kislinger, T. & Emili, A. Multidimensional protein identification technology: current status and future prospects. *Expert Rev. Proteomics* **2**, 27–39 (2005).
 30. Motoyama, A. & Yates, J. R. 3rd Multidimensional LC separations in shotgun proteomics. *Anal. Chem.* **80**, 7187–7193 (2008).
 31. Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D. A. & White, F. M. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl Acad. Sci. USA* **104**, 5860–5865 (2007).
 32. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428–432 (2009).
 33. George, T. C. *et al.* Distinguishing modes of cell death using the ImageStream multispectral imaging flow cytometer. *Cytometry A* **59**, 237–245 (2004).
 34. Edwards, B. S., Oprea, T., Prossnitz, E. R. & Sklar, L. A. Flow cytometry for high-throughput, high-content screening. *Curr. Opin. Chem. Biol.* **8**, 392–398 (2004).
 35. Krutzik, P. O. & Nolan, G. P. Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling. *Nature Methods* **3**, 361–368 (2006).
 36. Krutzik, P. O. & Nolan, G. P. Intracellular phospho-protein staining techniques for flow cytometry: monitoring single cell signaling events. *Cytometry A* **55**, 61–70 (2003).
 37. Nolan, G. P. Deeper insights into hematological oncology disorders via single-cell phospho-signaling analysis. *Hematology Am. Soc. Hematol. Educ. Program* 123–127, 509 (2006).
 38. Wei, G. *et al.* Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell* **10**, 331–342 (2006).
 39. Irish, J. M. *et al.* Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217–228 (2004).
 40. Krutzik, P. O., Crane, J. M., Clutter, M. R. & Nolan, G. P. High-content single-cell drug screening with phosphospecific flow cytometry. *Nature Chem. Biol.* **4**, 132–142 (2008).
- This work highlights a flow cytometry-based screen for inhibitors of multiple signalling pathways in heterogeneous primary cell populations at the single cell level.**
41. Lippincott-Schwartz, J., Snapp, E. & Kenworthy, A. Studying protein dynamics in living cells. *Nature Rev. Mol. Cell Biol.* **2**, 444–456 (2001).
 42. Giuliano, K. A., Haskins, J. R. & Taylor, D. L. Advances in high content screening for drug discovery. *Assay Drug Dev. Technol.* **1**, 565–577 (2003).
 43. Lang, P., Yeow, K., Nichols, A. & Scheer, A. Cellular imaging in drug discovery. *Nature Rev. Drug Discov.* **5**, 343–356 (2006).
 44. Venkatesh, N. *et al.* Chemical genetics to identify NFAT inhibitors: potential of targeting calcium mobilization in immunosuppression. *Proc. Natl Acad. Sci. USA* **101**, 8969–8974 (2004).
 45. Huang, K. & Murphy, R. F. From quantitative microscopy to automated image understanding. *J. Biomed. Opt.* **9**, 893–912 (2004).
 46. Abraham, V. C., Taylor, D. L. & Haskins, J. R. High content screening applied to large-scale cell biology. *Trends Biotechnol.* **22**, 15–22 (2004).
 47. MacDonald, M. L. *et al.* Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nature Chem. Biol.* **2**, 329–337 (2006).
 48. Perlman, Z. E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).
- This study describes cytological profiling of known drugs by a microscopy and data analysis method that successfully categorized blinded drugs.**
49. Tanaka, M. *et al.* An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol.* **3**, e128 (2005).
 50. Rabut, G. & Ellenberg, J. Automatic real-time three-dimensional cell tracking by fluorescence microscopy. *J. Microsc.* **216**, 131–137 (2004).
 51. Wang, M., Zhou, X., King, R. W. & Wong, S. T. Context based mixture model for cell phase identification in automated fluorescence microscopy. *BMC Bioinformatics* **8**, 32 (2007).
 52. Young, D. W. *et al.* Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nature Chem. Biol.* **4**, 59–68 (2008).
- In this study, factor analysis was used as a tool for rapid data reduction and to define cell phenotypes. Compound mechanisms of action were inferred from activity profiles integrated with predicted target binding profiles.**
53. Gao, Y. & Church, G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* **21**, 3970–3975 (2005).
 54. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
- This study describes a novel statistical method, the Connectivity Map, that uses gene expression profiles to find connections between small molecules that share a mechanism of action.**
55. Hieronymus, H. *et al.* Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10**, 321–330 (2006).
 56. Giuliano, K. A. *et al.* Systems cell biology knowledge created from high content screening. *Assay Drug Dev. Technol.* **3**, 501–514 (2005).
 57. Slack, M. D., Martinez, E. D., Wu, L. F. & Altschuler, S. J. Characterizing heterogeneous cellular responses to perturbations. *Proc. Natl Acad. Sci. USA* **105**, 19306–19311 (2008).
 58. Wang, J. *et al.* Cellular phenotype recognition for high-content RNA interference genome-wide screening. *J. Biomol. Screen* **13**, 29–39 (2008).
 59. Loo, L. H., Wu, L. F. & Altschuler, S. J. Image-based multivariate profiling of drug responses from single cells. *Nature Methods* **4**, 445–453 (2007).
- This paper describes a multivariate method to classify untreated and treated human cancer cells on the basis of single-cell phenotypic measurements. The classification provides a score, measuring the magnitude of the drug effect, and a vector, indicating the simultaneous phenotypic changes induced by the drug.**
60. Spearman, C. General intelligence, objectively determined and measured. *Am. J. Psychol.* **15**, 201–293 (1904).
 61. Jenkins, J. L., Bender, A. & Davies, J. W. *In silico* target fishing: predicting biological targets from chemical structure. *Drug Discov. Today Technol.* **3**, 413–421 (2006).
 62. Bender, A. *et al.* Using ligand based models for protein domains to predict novel molecular targets, and applications to triaging affinity chromatography data. *J. Proteome Res.* **8**, 2575–2585 (2009).
 63. Bender, A. *et al.* Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb. Chem. High Throughput Screen* **10**, 719–731 (2007).
 64. Prathipati, P., Ma, N., Manjunatha, U. & Bender, A. Fishing the target of antitubercular compounds: *in silico* target deconvolution model development and validation. *J. Proteome Res.* 20 Mar 2009 [doi:10.1021/pr8010843].
 65. Rines, D. R. *et al.* Whole genome functional analysis identifies novel components required for mitotic spindle integrity in human cells. *Genome Biol.* **9**, R44 (2008).
 66. Root, D. E., Hacohen, N., Hahn, W. C., Lander, E. S. & Sabatini, D. M. Genome-scale loss-of-function screening with a lentiviral RNAi library. *Nature Methods* **3**, 715–719 (2006).
 67. Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
- This study showed that phenotypic side-effect similarities can be used to infer whether two drugs share a target, and showed the feasibility of using phenotypic information to infer molecular interactions.**
68. Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960–1964 (2000).
 69. Nolan, G. P. What's wrong with drug screening today. *Nature Chem. Biol.* **3**, 187–191 (2007).
 70. Kitano, H. A robustness-based approach to systems-oriented drug design. *Nature Rev. Drug Discov.* **6**, 202–210 (2007).
 71. Fischbach, C. *et al.* Engineering tumors with 3D scaffolds. *Nature Methods* **4**, 855–860 (2007).
 72. Petersen, O. W., Ronnov-Jessen, L., Howlett, A. R. & Bissell, M. J. Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. *Proc. Natl Acad. Sci. USA* **89**, 9064–9068 (1992).
 73. Wichterle, H., Lieberam, I., Porter, J. A. & Jessell, T. M. Directed differentiation of embryonic stem cells into motor neurons. *Cell* **110**, 385–397 (2002).
 74. Torrance, C. J., Agrawal, V., Vogelstein, B. & Kinzler, K. W. Use of isogenic human cancer cells for high-throughput screening and drug discovery. *Nature Biotechnol.* **19**, 940–945 (2001).
 75. Janes, K. A. & Yaffe, M. B. Data-driven modelling of signal-transduction networks. *Nature Rev. Mol. Cell Biol.* **7**, 820–828 (2006).
 76. Sachs, K., Gifford, D., Jaakkola, T., Sorger, P. & Lauffenburger, D. A. Bayesian network approach to cell signaling pathway modeling. *Sci. STKE* PE38 (2002).
 77. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529 (2005).
 78. Chen, X., Zhou, X. & Wong, S. T. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE Trans. Biomed. Eng.* **53**, 762–766 (2006).
 79. Neumann, B. *et al.* High-throughput RNAi screening by time-lapse imaging of live human cells. *Nature Methods* **3**, 385–390 (2006).
 80. Faivre, S., Kroeber, G. & Raymond, E. Current development of mTOR inhibitors as anticancer agents. *Nature Rev. Drug Discov.* **5**, 671–688 (2006).
 81. Koehn, F. E. & Carter, G. T. The evolving role of natural products in drug discovery. *Nature Rev. Drug Discov.* **4**, 206–220 (2005).
 82. Terstappen, G. C., Schlupen, C., Raggiacchi, R. & Gaviraghi, G. Target deconvolution strategies in drug discovery. *Nature Rev. Drug Discov.* **6**, 891–903 (2007).