Review

CellPress

Integrating phenotypic small-molecule profiling and human genetics: the next phase in drug discovery

Cory M. Johannessen[1], Paul A. Clemons[2], and Bridget K. Wagner[2]

[1] Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA, USA
[2] Center for the Science of Therapeutics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

Over the past decade, tremendous progress in high-throughput small molecule-screening methods has facilitated the rapid expansion of phenotype-based data. Parallel advances in genomic characterization methods have complemented these efforts by providing a growing list of annotated cell line features. Together, these developments have paved the way for feature-based identification of novel, exploitable cellular dependencies, subsequently expanding our therapeutic toolkit in cancer and other diseases. Here, we provide an overview of the evolution of phenotypic small-molecule profiling and discuss the most significant and recent profiling and analytical efforts, their impact on the field, and their clinical ramifications. We additionally provide a perspective for future developments in phenotypic profiling efforts guided by genomic science.

Introduction

There are conceptual and logistical differences between high-throughput screening (HTS) and small-molecule profiling (see Glossary). A conventional HTS campaign typically deals with a single phenotypic readout under one set of conditions exposed to a large number of compounds. By contrast, profiling can involve multiple types of readout under similar conditions [1], binding profiles across many proteins [2,3], or, classically, a single readout across many cell lines. Several years ago, we reviewed the approach of small-molecule phenotypic profiling as a means of understanding the biological effects of chemical perturbation, as well as providing a blueprint for making chemical synthetic choices in constructing screening collections [4]. In subsequent years, an increasing stream of studies focused on profiling has infused the literature (Figure 1). A landmark study in the field described small-molecule profiling of the NCI-60 collection [5], in which cell growth measurements were made across 60 cell lines representing multiple cancer lineages, after perturbation with a common collection of nearly 4000 small molecules. Going a step further, the authors also profiled each cell line for protein expression of 76 common molecular targets in cancer; this combination of data was used to develop a matrix algebra-based correlation analysis of compounds with proteins. This study laid the foundation for, and foreshadowed, recent efforts to leverage genomic data during profiling [6–8]. Technical

Glossary

Bioactive compounds: small molecule perturbagens with previously known and annotated mechanism(s) of action
Cancer Cell Line Encyclopedia (CCLE): a project led by the Broad Institute, the Novartis Institute for Biomedical Research, and the Genomics Institute of the Novartis Research Foundation, focused on the genomic and pharmacologic characterization of a set of approximately 1000 human cancer cell lines (http://www.broadinstitute.org/ccle/home).
Cellular dependency: pathway or process on which a particular cell type or cell state depends for survival.
Chemical proteomics: measurement of the protein(s) to which a small molecule binds, directly or indirectly.
Gene-set enrichment analysis (GSEA): a computational method to determine the enrichment of a defined set of genes in the gene-expression analysis of two biological states. The collection of gene sets under consideration can be derived experimentally, computationally, or by biological function.
Interaction network: a graph-theoretic construct that expresses relations between proteins based on evidence of interaction between them (may be physical interaction, coregulation, or combinations of evidence).
Lead hopping: the identification of novel chemical matter with similar biological function, but different structure, to that of a starting small molecule.
Lymphoblastoid cell line (LCL): peripheral B lymphocytes transformed by EBV, these cells are often useful as a source of DNA and cells from individuals for biological characterization.
Matrix algebra: mathematical technique allowing connection of two variables via a third shared variable that connects to each one.
Meta-analysis: an analysis technique that merges data from separate studies of the same subject(s) into a single analysis.
Metabolomics: direct measurement of the levels of small-molecule metabolites in a cellular system under various conditions.
Molecular Libraries Program: a program operated by the National Institutes of Health focused on the discovery of novel chemical probes of biological processes and disease states (http://mli.nih.gov/mli/mlp-overview/)
Multifeature prediction: mathematical technique that combines information from multiple predictive variables to estimate the value of a predicted variable.
NCI-60 collection: a long-standing project operated by the National Cancer Institute involving the screening of a collection of 60 cancer cell lines for growth inhibition after treatment with small molecules of interest (http://dtp.nci.nih.gov/branches/btb/ivclsp.html).
Performance similarity: using performance of small molecules across multiple assays or using multiple features as a measure of their similarity (instead of, for example, chemical structure similarity).
Phenotypic pattern: a quantitative measurement of phenotype spanning multiple variables, such as gene-expression in response to compound treatment.
Quantitative trait locus (QTL): a genetic determinant of a quantitative trait of an organism, such as size.
Signature: specific set of multiple features reporting on a disease state or action of a perturbagen.
Small-molecule profiling: unbiased measurement of multiple cellular features reporting on small-molecule action.
Transcriptional profiling: unbiased measurement of multiple gene-expression features (usually as transcripts, possibly genome wide).

Corresponding author: Wagner, B.K. (bwagner@broadinstitute.org).
Keywords: cell line profiling; genotype–phenotype; small molecule.

0168-9525/
 http://dx.doi.org/10.1016/j.tig.2014.11.002

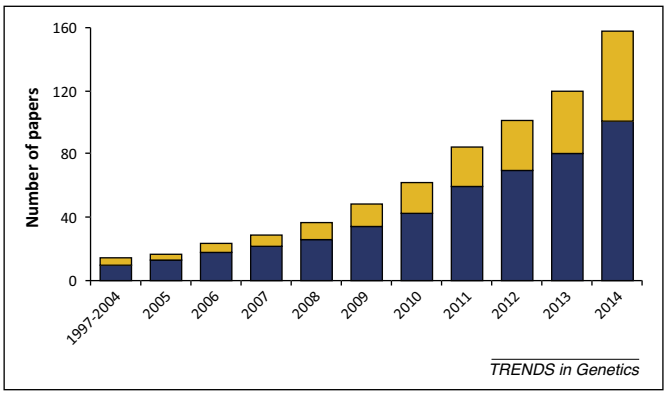

Figure 1. Cumulative increase in the number of papers in the literature focused on profiling experiments. Results were achieved by using the query 'phenotypic profiling', followed by manual annotation of each of the 100 results for their applicability to this metric. Papers were then labeled as specifically involving small-molecule profiling. We then added the literature cited in this review that did not use the query term to the total numbers. Yellow bars, papers involving a small molecule-based approach; blue bars, genetic perturbation or other phenotypic profiling.

advances in genomic sequencing and genotyping have enabled deeper characterization of cancer cell lines across multiple types of genetic feature, and high-throughput measurement of such cell lines subjected to hundreds or thousands of perturbations is now possible (Box 1).

The use of genome-wide features in evaluating profiling results means that relations that are more complex than simple single-gene associations may be uncovered. Of late, such genotype–phenotype correlations have been particularly important in cancer therapeutic discovery, because it has become increasingly clear that single drugs targeting cancer mutations may not be sufficient to treat all patients with a given tumor type. Improved diagnostics, coupled with precise small-molecule targeting, may ultimately fulfill the promise of 'personalized' medicine [9–11]. Here, we review recent efforts to leverage the explosion in genomic characterization data to exploit cellular dependencies for small-molecule sensitivity and resistance (Table 1). Although the cancer field has led the charge in this area, research focused on other complex genetic diseases can also benefit from the methods described here to arrive at more sophisticated models of disease and treatment.

Integrating profiling and human genetics: early efforts

Although the initial NCI-60 projects focused on determining the cancer lineage selectivity of compound response [5,12–15], researchers discovered, through early transcriptional profiling, that histopathological diagnoses of tumor types did not always correspond well with the gene-expression signatures [16]. Thus, it became clear that a more refined tumor characterization would ultimately be necessary to identify correlates of small-molecule sensitivity or resistance, and several efforts to do so soon followed. Transcriptional profiling of the NCI-60 collection across nearly 7000 genes, cross-analyzed with existing viability data, resulted in 232 gene expression-based signatures for compound sensitivity [17]. Again, classification was lineage independent, demonstrating the possibility that personalized approaches may be necessary for effective treatment, because therapies tailored to tumor type were insufficient in many cases. This idea is supported by a later study, in which close characterization of response to the mitogen-activated protein kinase kinase (MEK) inhibitor selumetinib showed that an 18-gene signature was sufficient to uncover activation of the MEK pathway, independent of lineage. This gene network served as a surrogate to measure MEK functional output; unexpectedly, this signature was not highly correlated with mutations in BRAF, RAS, or phosphatidylinositol 3-kinase (PI3K) [18]. Together, these studies helped drive home the point that a variety of genome-wide data (e.g., gene-expression data or mutational analysis) would be necessary to uncover vulnerabilities to small-molecule treatment.

Some efforts to couple profiling to human genetics have used Epstein–Barr virus (EBV)-transformed lymphoblastoid cell lines (LCLs) derived from human B lymphocytes. Given that these cell lines were used by the International HapMap Project to identify expression quantitative trait loci (QTLs) [19], they are densely genotyped and, thus, it was thought that this cell collection could provide the rich source of genetic annotation needed for small-molecule profiling. A test of 269 such LCLs against seven compounds with distinct mechanisms of toxicity provided data to model drug response in the context of RNA transcript levels and single nucleotide polymorphisms [20]. However, in this case, nongenetic confounders (e.g., cell growth rates or EBV levels used to transform the cells) prevented the study from reaching sufficient statistical power, suggesting that caution should be taken in the use of this cell source.

More disease-focused work also attempted to discover gene–small molecule interactions [21]. Using patient-derived LCLs from 18 members of a maturity-onset diabetes of the young 1 [MODY1, caused by mutation in the transcription factor hepatocyte nuclear factor 4 alpha (HNF4a)] family, the effects of nearly 4000 bioactive compounds and clinically used drugs were tested on these cells using an ATP-based viability readout. In a similar fashion to gene-set enrichment analysis (GSEA) [22], compound-set

Table 1. Phenotypic profiling efforts

Year	Area of focus	Cell lines	Compounds	Refs
1988–current	Growth inhibition	60	Ongoing	[5,12–15]
2008	Quantitative trait loci for compound sensitivity	269 LCLs	7	[20]
2011	MODY1	18 LCLs	~4000	[21]
2007	Kinase inhibitor profiling	500	14	[23]
2011	Kinase inhibitor profiling in Ewing's sarcoma	2	200	[24]
2009	Genomic characterization of nonsmall cell lung cancer	84	12	[28]
2012	Wide-ranging cancer cell line characterization	639	130	[6]
2012	Wide-ranging cancer cell line characterization	949	24	[7]
2013	Expansion of small molecules tested	242	354	[8]

enrichment analysis of these data revealed classes of small molecule connected with HNF4α mutation, including a subset of fatty acids. Remarkably, many of these compounds retained activity in pancreatic beta cells, enhancing glucose-stimulated insulin secretion in cells in which HNF4α had been knocked down. The authors focused on the utility of this approach in determining the functional-consequence of particular disease alleles, but this work also helped lay the groundwork for profiling large numbers of genetically characterized cancer cell lines to identify genetic bases for compound sensitivity or resistance, and to uncover previously unknown cancer dependencies.

Establishing a profiling matrix

Studies during the 2000s tended to focus on one 'face' of the profiling matrix (Figure 2), such as a particular compound class or specific types of cancer cell line. In the former case, 500 cancer cell lines were profiled against 14 kinase inhibitors with known targets and well-annotated mechanisms of action [23]. The particular kinases targeted were examined in these cell lines for relations between inhibitor effects and mutational status, and known relations were confirmed, foreshadowing larger-scale efforts to come. Although many of the cellular sensitivities uncovered by this approach recapitulated known disease genotype–phenotype connections, they highlighted the existence of low-frequency, drug response-associated genetic events in tumor types not broadly considered to be drug sensitive. These studies reinforced the notion that, at least for this particular class of compounds, the underlying genetic events are a stronger determinant of sensitivity and response than are disease subtype, lineage, or tissue of origin. By contrast, a recent effort to profile two Ewing's sarcoma cell lines against 200 kinase inhibitors [24] could be thought of as the transpose of such a matrix, where in this case only a few cell lines were profiled against many compounds. This study revealed that tozasertib had unique and selective efficacy against these cells; chemical proteomics identified inhibition of both Aurora A and B as being responsible for this activity. However, this outcome may be the result of examining a narrow biology and small number of cell lines and may not be borne out by broader experiments.

Alongside these more unbiased methods, approaches designed to look at particular genetic or small-molecule hypotheses were developed and applied to small-molecule profiling. For example, it had been proposed that RAS and BRAF mutations would be similarly sensitive to MEK inhibition, because mutations are often mutually exclusive, and impinge upon similar downstream signaling cascades. However, the discovery that cells harboring the BRAF mutation V600E, regardless of lineage, were selectively sensitive to MEK inhibitors over cells harboring RAS mutations [25] lent substantial credit to the promise of personalized medicine, and led to the development of vemurafenib, a potent MEK inhibitor used for treating malignant melanoma [26,27].

Moving toward full integration: expansion of genetic features

Previous studies successfully demonstrated that phenotypic profiling had great power to identify compounds with unique sensitivity profiles across cancer cell lines. However, it was clear that the full potential of this approach would only be realized when linked to additional cellular

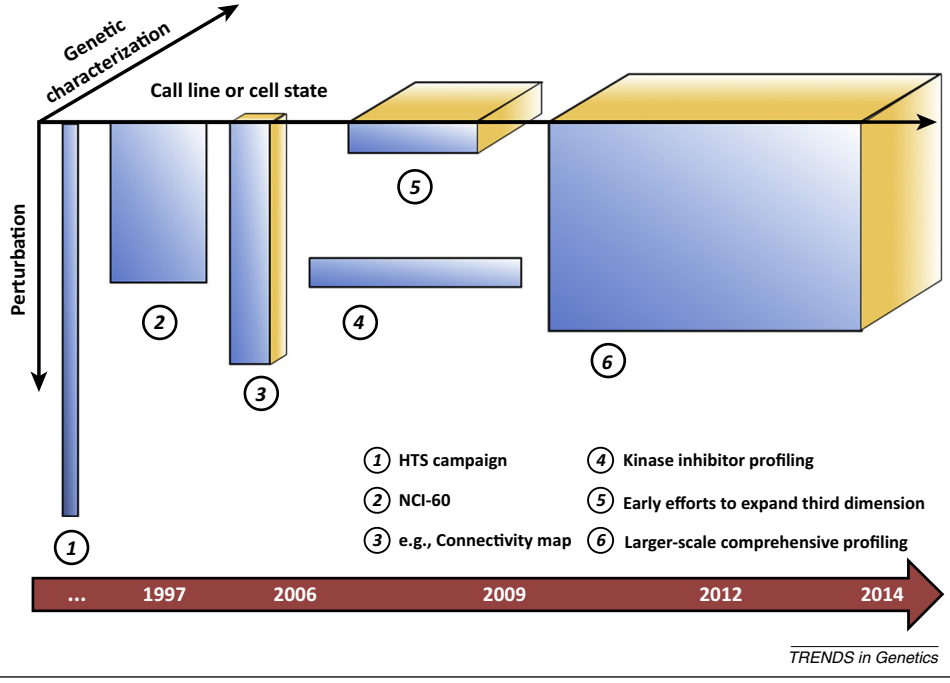

Figure 2. The evolution of a profiling matrix over the course of the past 20 years. The three axes represent perturbations (small molecule or genetic), cell lines or cell states profiled, and genetic features characterized. **(1)** A typical high-throughput screening (HTS) campaign; **(2)** experimentation exemplified by the NCI-60 collection [5]; **(3)** the inclusion of gene-expression data in profiling experiments [72,73]; **(4)** profiling of a few related compounds across many cell lines [21]; **(5)** early attempts to expand the genetic features characterized across profiled cell lines [26]; **(6)** efforts to expand the matrix to incorporate many cell lines, many genetic features, and now many compounds [6–8].

features. Historically, performing small-molecule screens on characterized cell line models required a series of compromises, either in cell line, genomic feature, or small-molecule space. Although the potential of a truly comprehensive effort was clear, the financial and logistical challenges required represented a formidable challenge. To address this challenge, a panel of 84 nonsmall cell lung cancer cell lines was profiled against 12 compounds under clinical evaluation [28]. A key difference in this study was the high degree of genomic characterization of the cell lines, including analyses of copy number variation and oncogenic mutation in addition to gene-expression profiling. Importantly, this additional characterization revealed the genomic similarity between cancer cell lines and primary tumors, making possible the large-scale profiling for which cell lines are required. The authors found that KRAS mutations sensitized cells to heat shock protein (Hsp)90 inhibition, whereas increased copy numbers of ABL2 and SRC induced sensitivity to the SRC/ABL inhibitor dasatinib. Such observations would probably be missed with only gene expression-profiling data.

This and other studies highlighted the need to maximize cellular models and associated cellular features (genomic or otherwise) to fully interpret the phenotypic patterns elicited by small-molecule collections. Recently, two parallel efforts from the Sanger Institute/Massachusetts General Hospital (MGH) and the Broad Institute/Novartis have greatly expanded the matrix, assembling collections of 639 (Sanger/MGH) and 949 (Broad/Novartis) individual cancer cell lines [6,7]. These collections were deeply characterized for mutations in known cancer genes, whole-transcriptome mRNA expression, DNA copy number alterations, and sensitivity to a panel of small molecules (130 in the Sanger study, 24 in the Broad study). Collectively, these efforts have had a profound impact on the ability to discern genetic markers of drug sensitivity and resistance within these collections, including RAF/MEK inhibitors in melanoma and bromodomain inhibitors in N-Myc-amplified neuroblastoma [29,30].

Cancer cell line profiling studies, especially the analysis of such rich data sets, are still in a developmental phase. The Sanger/MGH [6] and Broad/Novartis [7] studies used different assay technologies in their profiling experiments and subsequently different data-analysis methods, including distinct methods of summarizing cell line sensitivity data and differential emphases on single- versus multi-feature prediction of important genetic features. Such differences in methods, both experimental and analytical, have drawn criticism from some quarters. For example, in a meta-analysis of these two studies [31], the authors explicitly compared the concordance of both the gene-expression and pharmacological assay measurements that underlie the two studies, focusing particularly on 15 small molecules tested in both studies. The authors noted high concordance among the gene-expression measurements between the studies, but relatively low concordance between the sensitivity measurements. They interpreted these inconsistencies to mean that all of the measured pharmacological responses of both studies were questionable, because it was not clear from the available data which study provided the most meaningful assay.

A more nuanced view [32] suggested that differences between the two studies might be explained by any combination of differences in expression profiles, sensitivity measurements, or computational methods. In particular, it was pointed out that many differences in the experimental methods between the studies that might be expected to influence the results, including different assays, compound concentrations, and cell culture conditions. Clearly, cancer cell line profiling with small molecules requires maturation both in data collection and analysis procedures (relative to, say, the more mature gene-expression profiling). However, it is precisely by performing these large-scale experiments that the research community will learn which methods are more likely to succeed in uncovering new cancer dependencies. As such, more attention should be focused on explaining discrepancies between these studies in terms of the real, technical differences between them.

Similar to the model advanced by the NCI through the NCI-60 collection, these data sets are not static: they are modular entities that can be expanded in the future (Figure 2). For example, the Broad/Novartis Cancer Cell Line Encyclopedia (CCLE) provides deep genomic characterization of the ~950 cell lines in the collection, but was tested against relatively few (24) compounds. It was clear that a larger collection of small molecules needed to be profiled to maximize the power of this approach. To that end, Basu *et al.* acquired a subset of 242 CCLE cancer cell line models and profiled them against an expanded collection of 354 small molecules, selected for their high selectivity and targeting of many proteins known or thought to be hubs of cell signaling [8]. In addition to expanding the overarching chemical biology coverage of this data set, this study highlighted an unappreciated dependency on anti-apoptotic Bcl-2 family members in the context of activating β-catenin mutations. Collectively, these efforts suggest that our discovery capabilities are not saturated, and that enhanced cell models and compound collections are likely to further our ability to identify new therapeutic hypotheses and to refine those nominated in earlier studies.

Importantly, Basu *et al.* [8] also explicitly exploited cell lineage or tissue of origin as a variable in their analysis, and highlighted the importance of confounding factors, such as whether cells were grown in suspension or were adherent to the tissue culture substrate. This study focused primarily on the association of mutation and lineage features of CCLs with response to small molecules, resulting in a large number of such connections and a public resource that users can query using compounds or genes of interest [8] (http://broadinstitute.org/ctrp). A more recent study from the same group systematically examined correlations of small-molecule sensitivity for 481 compounds with genome-wide basal gene expression among 860 CCLs. Such analyses revealed unique and diverse insights into drug metabolism and mechanism of action in cancer cells, including identification of new target pathways and partner proteins (M. Rees *et al.*, 2014 unpublished). Although the sensitivity measurements used were only correlated with gene-expression data, by amassing enough data across a large number of cell lines, the authors were able to discern significant patterns of connecting expression to compound sensitivity.

Using multigene features to find genomic correlates of chemical sensitivity

With the advent of large pharmacological data sets, finding genomic correlates of drug sensitivity will increasingly need to leverage the predictive power of using sets of multiple genomic features. The discoveries made using single-gene approaches likely represent only the tip of the iceberg, and we are far from saturation. The most obvious multigene approaches rely heavily on the body of existing knowledge, such as pathway databases and protein–protein interaction networks. Methods such as GSEA [22] offer a starting point, although their utility in analyzing these experiments depends heavily on the choice of gene-set collection used, which must be specified in advance [33]. One study proposes an alternative gene set collection, based on co-expressed genes in cancer, to those available in GSEA as superior for cancer cell line-profiling data, because more of the gene sets in this collection were significantly enriched after treatment of cancer cell lines with a collection of small molecules [33]. Indeed, many methods rely on the state and completeness of sources of prior knowledge, because such data may be incomplete. The relevance of such prior knowledge is itself a subject of study, in which network biology models having or lacking such prior information are directly compared [34]. Unbiased methods, such as elastic net and related regression techniques [35], have been used to discover cancer dependencies [7,8], but their application to these data had not been systematically evaluated until recently [36]. In general, models emerging from multigene-predictive methods still require biological interpretation of the lists of genes that represent model features. Most such analyses tend to center on genes recognizable by name to cancer biologists.

One way to analyze cancer cell line profiling to exploit dependencies based on multiple features is to create a signature of a particular cell state (e.g., glycolytic versus respiring [37,38]) and correlate this signature with sensitivity to small molecules. Cell state signatures are in principle accessible using any 'omic data, but most commonly from gene-expression data. In a recent example, Viswanathan *et al.* used gene-expression and pathway analysis to prepare a signature of the progression of cell lines relative to the epithelial-to-mesenchymal transition (EMT) (V. Viswanathan et al., 2014, unpublished). By correlating this signature with small-molecule sensitivity data, these authors showed that inhibitors of glutathione peroxidase 4 (GPX4) exhibited selectivity for epithelium-derived cancer cells that had undergone EMT, as well as for cancer cells of mesenchymal origin.

Another recent study explored the associations between groups of small molecules targeting common proteins or pathways and groups of cell lines sharing cellular or genetic features (B. Seashore-Ludlow *et al.*, 2014, unpublished). Clustering sensitivity data, followed by mining clusters for enriched hot-spots, allows unbiased identification of clusters of CCLs sharing common genetic features that respond similarly to small molecules with shared targets. These studies provide a blueprint for moving toward profiling analyses that incorporate a greater proportion of overall genomic data to identify correlates of small-molecule sensitivity.

Identifying resistance mechanisms through profiling

The identification of small-molecule dependencies in cancer has had a direct and profound impact on clinical outcomes. Despite this progress, the long-term efficacy of anticancer therapies, including both targeted and cytotoxic agents, is almost universally thwarted by the acquisition of drug resistance, representing an unmet clinical challenge. Moreover, as our collection of therapeutic modalities expands, this challenge is likely to continue to grow in scope and size, increasing the necessity for understanding the genetic underpinnings of therapeutic resistance. For some time, drug resistance appeared to be a complex and insurmountable phenomenon. However, the advent of systematic methods of identifying resistance mechanisms has, similar to pharmacogenomic screening approaches, enabled the community to start defining the principles of therapeutic resistance. Much of this progress has leveraged the development of comprehensive genetic perturbation libraries, including gain- [cDNA/open reading frame (ORF)] and loss-of-function [short hairpin (sh)RNA/clustered regularly interspaced short palindromic repeats (CRISPR)] reagents, but the principles followed are the same for small-molecule profiling. These reagents have successfully been used as a systematic and comprehensive means of rescuing the phenotype associated with small-molecule dependencies. An initial demonstration of this approach used a genome-scale shRNA library to identify loss-of-function events that drive therapeutic resistance to imatinib in a cell line model of CML, leading to identification of protein tyrosine phosphatase, non-receptor type 1 (*PTPN1*), neurofibromin 1 (*NF1*), SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily b, member 1 (*SMARCB1*), and *SMARCE1* as genes essential for imatinib response [39].

More recently, cDNA (or ORF) gene-expression libraries have caught up with loss-of-function libraries, enabling screening for genes sufficient to induce therapeutic resistance [40]. The initial iteration of this experiment involved the screening of a kinase-centric library, which was successfully used to identify COT/mitogen-activated protein kinase kinase kinase 8 (MAP3K8) as a driver of resistance to RAF inhibitors in BRAFV600E-mutant melanoma. Subsequent efforts using this screening resource identified several kinase mediators of resistance, including p90RSK3/4 (PI3K inhibition [41]), Crk-like protein {CRKL; epidermal growth factor receptor (EGFR) inhibition [42]} and protein kinase A [PKA; human epidermal growth factor 2 (HER2) inhibition [43]]. Recently, this approach has been expanded beyond the kinome to a near-genome scale collection of cDNAs. Using malignant melanoma as a model system for understanding therapeutic resistance, Johannessen *et al.* [44] screened this collection for genes that mediate resistance to RAF, MEK, extracellular signal-regulated kinase (ERK), and combined RAF/MEK inhibition in the context of the $BRAF^{V600E}$ mutation. These studies revealed a complex and diverse set of kinase and nonkinase resistance effectors, including transcription factors, signaling pathways operant outside of the RAF/MEK/ERK pathway, and a series of other uncharacterized genes. In addition to loss-of-function and wildtype ORF libraries, mutational scanning, for example of kinase drug targets [45–48], has been a particularly useful tool for

identifying resistance mutations. Recent technologies enabling the creation of fully saturated mutational scanning libraries are likely to further increase our understanding of drug–target interactions, as well as annotating resistance alleles [49]. Together, these studies advance the notion that large-scale profiling experiments can uncover genetic targets for therapeutic intervention.

Concluding remarks and outlook

A key feature enabling the success of a genomic approach to phenotypic drug discovery has been the detailed level of characterization of cancer cell lines, and the fact that such models have been shown to faithfully recapitulate the genetics and small-molecule sensitivity of primary patient samples [25,50]. However, to expand beyond cancer, disease-focused communities need to create similar tools and models to enable modernization of drug discovery in their disease areas. A recent example is provided by a chemical genomics approach taken toward malaria [51]. In that case, 2816 compounds were profiled in eight-point concentration-response against 61 parasite lines, and the results were combined with genome-wide association studies and linkage analysis for selective sensitivity. In particular, compounds with differential chemical phenotypes across three parental parasite strains were investigated more fully. Remarkably, 96% of these differences were mapped to three loci containing genes encoding dihydrofolate reductase, multidrug resistance protein, and chloroquine resistance transporter. One outcome resulting from such a data set was a set of novel strategies for overcoming drug resistance in malaria, but the point that stands out here is the high degree of analogy to cancer cell line profiling. In both cases, the goal is to find genetic determinants of susceptibility to small-molecule perturbation. Furthermore, the similarity between resistance in cancer and infectious disease has been recognized [52].

Exploring beyond the DNA-encoded genome for cellular vulnerabilities to small-molecule perturbation is likely to be an important next direction for phenotypic profiling. For example, epigenetic modifications represent reversible changes that might be leveraged in designing profiling experiments [53], and several chromatin-modifying enzymes are under heavy scrutiny, such as enhancer of zeste homolog 2 (EZH2) and DOT1-like histone H3K79 methyltransferase (DOT1L) [54–56]. Similarly, as proteomics and metabolomics techniques become more miniaturized and routine, profiling the functional genome [57,58] or metabolome [59,60] for cancer cell dependencies will become more of a reality. Analysis methods, such as similarity network fusion [61], optimized to take advantage of such integrated data sets, are already in development. Such methods will be particularly important as we move toward larger data sets, for example those that use genomic measurements following each perturbation, as opposed to in the steady state.

The development of disease-relevant readouts and models is essential for the success of linking genotype to phenotype. Again, this approach has been most evident in the cancer community, where the NCI-60 collection has had so much influence [5,15]. More recently, deliberate efforts to create panels of cell lines have enabled greater statistical power across lineages and mutational status [62]. Lymphoblastoid cell lines, for example, derived from the International HapMap project [63], may have their own shortcomings regarding reproducibility and other confounding factors [20]. Ultimately, the use of techniques to generate iPS cells from patient samples, followed by directed differentiation to relevant cell types [64–66], may be the most productive avenue for the future. It remains to be seen whether these approaches can bear fruit in cases where cell death is not the desired phenotype, but the establishment of an NCI-60-like collection for other diseases would go a long way toward answering this question.

Box 1. Recent advances in small-molecule profiling

A method of small-molecule profiling that has begun to be explored is building profiles using historical screening data. Such profiling is most useful for screening collections exposed together to many assays over time, usually in the same facility, so that a large number of measurements accumulate. Even though the assays were individually conceived and not intended as a profile, it is possible to normalize or discretize such data to make meaningful calculations of performance similarity between compounds. In one recent study [67], computational scientists profiled historical screening data across 195 assays run on approximately 1.5 million compounds. The authors show applications for their method in developing hypotheses about mechanism of action (MoA), lead hopping, and selection of library subsets. A similar study profiled historical screening data [68] and likewise showed the ability of such data to generate MoA hypotheses and to group compounds into communities based solely on their performance. Whenever a common compound collection is exposed to many such assays, these methods hold great promise, a fact recognized by the National Institutes of Health when it commissioned a new public interface, the BioAssay Research Database (BARD), to showcase the findings of the Molecular Libraries Program [69] (E. Howe *et al.*, 2014, unpublished).

Although powerful, methods that need accumulation of data are inherently passive: they require that data sets be allowed to accumulate over extended periods of time. In each of the above cases, approximately 10 years of data collection were performed before the data-mining methods were published and presented. A more active approach to profiling is to expose small-molecule collections prospectively to high-dimensional multiplex measurements; recent advances in assay technology make this possible. High-content screening, usually involving high-throughput microscopy and image analysis, has often been at the forefront of method development [70–72]. A recently developed imaging assay using multiple fluorescent dyes and channels was used for unbiased cell morphological profiling of small-molecule action [73], analogous to the more familiar gene-expression profiling of small molecules [74,75]. In general, high-dimensional data collection per well and hierarchical clustering methods have been used to study MoA for novel compounds [75,76], and can also be used to develop profile-based structure-activity relations among molecules [77]. In one recent study, a large collection of compounds was exposed to profiling by both multiplex imaging and gene-expression methods, and the resulting data were used to select a performance-diverse subset of compounds [76]. These technological advances reduce some of the barriers to assembling informer sets for small-molecule profiling, freeing up time and effort to devote to more sophisticated data analysis and integration.

Acknowledgments

The preparation of this manuscript was supported by a Career Development Award from the Melanoma Research Foundation (C.M.J), the Cancer Target Discovery and Development Network (NCI, RC2-CA148399), and the JDRF (B.K.W.).

References

1 Wagner, B.K. *et al.* (2008) Large-scale chemical dissection of mitochondrial function. *Nat. Biotechnol.* 26, 343–351
2 Clemons, P.A. *et al.* (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18787–18792
3 Kauvar, L.M. *et al.* (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* 2, 107–118
4 Wagner, B.K. and Clemons, P.A. (2009) Connecting synthetic chemistry decisions to cell and genome biology using small-molecule phenotypic profiling. *Curr. Opin. Chem. Biol.* 13, 539–548
5 Weinstein, J.N. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349
6 Garnett, M.J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483, 570–575
7 Barretina, J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607
8 Basu, A. *et al.* (2013) An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 154, 1151–1161
9 Ruden, D.M. (2007) Personalized medicine and quantitative trait transcripts. *Nat. Genet.* 39, 144–145
10 van't Veer, L.J. and Bernards, R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452, 564–570
11 Beckman, R.A. *et al.* (2012) Impact of genetic dynamics and single-cell heterogeneity on development of nonstandard personalized medicine strategies for cancer. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14586–14591
12 Paull, K.D. *et al.* (1992) Identification of novel antimitotic agents acting at the tubulin level by computer-assisted evaluation of differential cytotoxicity data. *Cancer Res.* 52, 3892–3900
13 Paull, K.D. *et al.* (1989) Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl. Cancer Inst.* 81, 1088–1092
14 Shoemaker, R.H. *et al.* (1988) Development of human tumor cell line panels for use in disease-oriented drug screening. *Prog. Clin. Biol. Res.* 276, 265–286
15 Shoemaker, R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* 6, 813–823
16 Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537
17 Staunton, J.E. *et al.* (2001) Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10787–10792
18 Dry, J.R. *et al.* (2010) Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). *Cancer Res.* 70, 2264–2273
19 International HapMap, C. (2005) A haplotype map of the human genome. *Nature* 437, 1299–1320
20 Choy, E. *et al.* (2008) Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* 4, e1000287
21 Shaw, S.Y. *et al.* (2011) Disease allele-dependent small-molecule sensitivities in blood cells from monogenic diabetes. *Proc. Natl. Acad. Sci. U.S.A.* 108, 492–497
22 Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550
23 McDermott, U. *et al.* (2007) Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19936–19941
24 Winter, G.E. *et al.* (2011) An integrated chemical biology approach identifies specific vulnerability of Ewing's sarcoma to combined inhibition of Aurora kinases A and B. *Mol. Cancer Ther.* 10, 1846–1856
25 Solit, D.B. *et al.* (2006) BRAF mutation predicts sensitivity to MEK inhibition. *Nature* 439, 358–362
26 Bollag, G. *et al.* (2010) Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. *Nature* 467, 596–599
27 Flaherty, K.T. *et al.* (2012) Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. *N. Engl. J. Med.* 367, 1694–1703
28 Sos, M.L. *et al.* (2009) Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. *J. Clin. Invest.* 119, 1727–1740
29 Puissant, A. *et al.* (2013) Targeting MYCN in neuroblastoma by BET bromodomain inhibition. *Cancer Discov.* 3, 308–323
30 Konieczkowski, D.J. *et al.* (2014) A melanoma cell state distinction influences sensitivity to MAPK pathway inhibitors. *Cancer Discov.* 4, 816–827
31 Haibe-Kains, B. *et al.* (2013) Inconsistency in large pharmacogenomic studies. *Nature* 504, 389–393
32 Weinstein, J.N. and Lorenzi, P.L. (2013) Cancer: discrepancies in drug sensitivity. *Nature* 504, 381–383
33 Bateman, A.R. *et al.* (2014) Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci. Rep.* 4, 4092
34 Olsen, C. *et al.* (2014) Relevance of different prior knowledge sources for inferring gene interaction networks. *Front. Genet.* 5, 177
35 Zhu, J. and Hastie, T. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics* 5, 427–443
36 Jang, I.S. *et al.* (2014) Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac. Symp. Biocomput.* 2014, 63–74
37 Ramanathan, A. *et al.* (2005) Perturbational profiling of a cell-line model of tumorigenesis by using metabolic measurements. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5992–5997
38 Gohil, V.M. *et al.* (2010) Nutrient-sensitized screening for drugs that shift energy metabolism from mitochondrial respiration to glycolysis. *Nat. Biotechnol.* 28, 249–255
39 Luo, B. *et al.* (2008) Highly parallel identification of essential genes in cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20380–20385
40 Yang, X. *et al.* (2011) A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* 8, 659–661
41 Serra, V. *et al.* (2013) RSK3/4 mediate resistance to PI3K pathway inhibitors in breast cancer. *J. Clin. Invest.* 123, 2551–2563
42 Cheung, H.W. *et al.* (2011) Amplification of CRKL induces transformation and epidermal growth factor receptor inhibitor resistance in human non-small cell lung cancers. *Cancer Discov.* 1, 608–625
43 Moody, S.E. *et al.* (2014) PRKACA mediates resistance to HER2-targeted therapy in breast cancer cells and restores anti-apoptotic signaling. *Oncogene* Published online June 9, 2014, (http://dx.doi.org/10.1038/onc.2014.153)
44 Johannessen, C.M. *et al.* (2013) A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. *Nature* 504, 138–142
45 Azam, M. *et al.* (2003) Mechanisms of autoinhibition and STI-571/imatinib resistance revealed by mutagenesis of BCR-ABL. *Cell* 112, 831–843
46 Emery, C.M. *et al.* (2009) MEK1 mutations confer resistance to MEK and B-RAF inhibition. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20411–20416
47 Wagle, N. *et al.* (2011) Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *J. Clin. Oncol.* 29, 3085–3096
48 Antony, R. *et al.* (2013) C-RAF mutations confer resistance to RAF inhibitors. *Cancer Res.* 73, 4840–4851
49 Melnikov, A. *et al.* (2014) Comprehensive mutational scanning of a kinase in vivo reveals substrate–dependent fitness landscapes. *Nucleic Acids Res.* 42, e112
50 Neve, R.M. *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10, 515–527
51 Yuan, J. *et al.* (2011) Chemical genomic profiling for antimalarial therapies, response signatures, and molecular targets. *Science* 333, 724–729
52 Glickman, M.S. and Sawyers, C.L. (2012) Converting cancer therapies into cures: lessons from infectious diseases. *Cell* 148, 1089–1098
53 Mair, B. *et al.* (2014) Exploiting epigenetic vulnerabilities for cancer therapeutics. *Trends Pharmacol. Sci.* 35, 136–145
54 McCabe, M.T. *et al.* (2012) EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature* 492, 108–112

55 Daigle, S.R. *et al.* (2013) Potent inhibition of DOT1L as treatment of MLL-fusion leukemia. *Blood* 122, 1017–1025
56 Wee, Z.N. *et al.* (2014) EZH2-mediated inactivation of IFN-gamma-JAK-STAT1 signaling is an effective therapeutic target in MYC-driven prostate cancer. *Cell Rep.* 8, 204–216
57 Fagerberg, L. *et al.* (2011) Large-scale protein profiling in human cell lines using antibody-based proteomics. *J. Proteome Res.* 10, 4066–4075
58 Pernemalm, M. *et al.* (2013) Quantitative proteomics profiling of primary lung adenocarcinoma tumors reveals functional perturbations in tumor metabolism. *J. Proteome Res.* 12, 3934–3943
59 Locasale, J.W. *et al.* (2011) Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. *Nat. Genet.* 43, 869–874
60 Brunelli, L. *et al.* (2014) Capturing the metabolomic diversity of KRAS mutants in non-small-cell lung cancer cells. *Oncotarget* 5, 4722–4731
61 Wang, B. *et al.* (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337
62 Sharma, S.V. *et al.* (2010) Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat. Rev. Cancer* 10, 241–253
63 International HapMap, C. (2003) The International HapMap Project. *Nature* 426, 789–796
64 Brennand, K.J. *et al.* (2011) Modelling schizophrenia using human induced pluripotent stem cells. *Nature* 473, 221–225
65 Israel, M.A. *et al.* (2012) Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. *Nature* 482, 216–220
66 Gafni, O. *et al.* (2013) Derivation of novel human ground state naive pluripotent stem cells. *Nature* 504, 282–286
67 Petrone, P.M. *et al.* (2012) Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem. Biol.* 7, 1399–1409
68 Dancik, V. *et al.* (2014) Connecting small molecules with similar assay performance profiles leads to new biological hypotheses. *J. Biomol. Screen.* 19, 771–781
69 de Souza, A. *et al.* (2014) An overview of the challenges in designing, integrating, and delivering BARD: a public chemical-biology resource and query portal for multiple organizations, locations, and disciplines. *J. Biomol. Screen.* 19, 614–627
70 Young, D.W. *et al.* (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* 4, 59–68
71 Futamura, Y. *et al.* (2012) Morphobase, an encyclopedic cell morphology database, and its use for drug target identification. *Chem. Biol.* 19, 1620–1630
72 Ljosa, V. *et al.* (2013) Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.* 18, 1321–1329
73 Gustafsdottir, S.M. *et al.* (2013) Multiplex cytological profiling assay to measure diverse cellular states. *PLoS ONE* 8, e80999
74 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
75 Lamb, J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935
76 Wawer, M.J. *et al.* (2014) Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10911–10916
77 Wawer, M.J. *et al.* (2014) Automated structure-activity relationship mining: connecting chemical structure to biological profiles. *J. Biomol. Screen.* 19, 738–748