The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease

Justin Lamb,[1]* Emily D. Crawford,[1]† David Peck,[1] Joshua W. Modell,[1] Irene C. Blat,[1] Matthew J. Wrobel,[1] Jim Lerner,[1] Jean-Philippe Brunet,[1] Aravind Subramanian,[1] Kenneth N. Ross,[1] Michael Reich,[1] Haley Hieronymus,[1,2] Guo Wei,[1,2] Scott A. Armstrong,[2,3] Stephen J. Haggarty,[1,4] Paul A. Clemons,[1] Ru Wei,[1] Steven A. Carr,[1] Eric S. Lander,[1,5,6] Todd R. Golub[1,2,3,5,7]*

To pursue a systematic approach to the discovery of functional connections among diseases, genetic perturbation, and drug action, we have created the first installment of a reference collection of gene-expression profiles from cultured human cells treated with bioactive small molecules, together with pattern-matching software to mine these data. We demonstrate that this "Connectivity Map" resource can be used to find connections among small molecules sharing a mechanism of action, chemicals and physiological processes, and diseases and drugs. These results indicate the feasibility of the approach and suggest the value of a large-scale community Connectivity Map project.

A fundamental challenge that arises throughout biomedicine is the need to establish the relation among diseases, physiological processes, and the action of small-molecule therapeutics. Our goal is to provide a generic solution to this problem by attempting to describe all biological states—physiological, disease, or induced with a chemical or genetic construct—in terms of genomic signatures, create a large public database of signatures of drugs and genes, and develop pattern-matching tools to detect similarities among these signatures. Using such a resource, a researcher studying a drug candidate, a gene, or a disease state could compare its signature to the database to discover unexpected connections—much as one can compare a DNA sequence to the GenBank database to identify similar genes. We will refer to this resource as a "Connectivity Map" because of its potential to reveal "connections" among drugs, genes, and diseases.

In principle, there are many possible genomic signatures that might be used—including DNA methylation patterns, mRNA levels, and protein expression or metabolite profiles. To be practical, however, such signatures should be generated from a small number of cells at low cost, in high throughput, and with sufficiently high complexity to provide a rich description. At present, only mRNA expression assayed on DNA microarrays meets these criteria. We have therefore chosen this as the "universal language" with which to describe cellular responses.

Gene-expression profiling has historically been applied in specific settings to elucidate the mechanisms underlying a biological pathway (*1*, *2*), to reveal cryptic subtypes of a disease (*3*, *4*), and to predict cancer prognosis (*5*, *6*). But here we envisage its use as the means to catalog the biological responses to a large number of diverse perturbations. Of course, this idea is not entirely new. A landmark study by Hughes *et al.* (*7*) demonstrated that a compendium of gene-expression data could be used for the functional annotation of small molecules and genes, at least in yeast. Although that study was encouraging, the extent to which the approach would be applicable to mammalian biology was not obvious. More recently, a variety of commercial databases of expression profiles from rat tissues after systemic administration of known drugs have been developed [e.g., (*8*)], and these appear to have value for the identification of potential toxicities of new chemical entities [e.g., (*9*)]. However, such in vivo studies suffer from serious practical limitations. First, the type of perturbagens that can be studied is limited. Only small molecules with druglike physicochemical properties can be effectively administered to live animals. And systematic genetic perturbation (i.e., with RNA interference) is not yet possible. Second, the high cost of whole-animal studies precludes contemplating such an approach at the genome scale.

We hypothesized that perturbations in mammalian cell culture might provide an approach that is truly generalizable, systematic, and biologically relevant. However, several potential pitfalls must be considered. Conceivably, a large number of parameters would need to be optimized for each perturbation, including cell type, concentration, and treatment duration. Equally, analytical methods capable of detecting relevant signals in the data might not be generally applicable. If so, generation of a useful Connectivity Map would be impractical. However, here we demonstrate—through the recovery of known, and the discovery of new, biological connections—that the Connectivity Map concept is indeed viable.

Creating a First-Generation Connectivity Map

Perturbagens. We studied 164 distinct small-molecule perturbagens, selected to represent a broad range of activities, and including U.S. Food and Drug Administration (FDA)–approved drugs and nondrug bioactive "tool" compounds. We included multiple compounds sharing molecular targets (e.g., histone deacetylase inhibitors) to determine whether such compounds would share a molecular signature. Similarly, we profiled compounds with the same clinical indication (e.g., antidiabetics), which allowed us to determine whether connections could be established on the basis of therapeutic class, even though the mechanisms of action might be distinct. Furthermore, we chose some small molecules that act proximal to gene expression (e.g., selective estrogen receptor modulators) and some whose primary targets are much more distal (e.g., immunomodulators, inhibitors of signal transduction). Finally, we included some compounds whose targets are not expressed in all cell types (e.g., COX2 inhibitors), whose clinical effects are non–cell-autonomous (e.g., aromatase inhibitors), or whose activities are only discernible after chronic, in vivo exposure (e.g., antipsychotics).

Cell lines. Ideally, one would generate profiles in a wide diversity of established and primary cells, but practicality limits us to only a few lines that can be stably grown over long periods of time. For this pilot study we generated most of our data in the breast cancer epithelial cell line MCF7 because it has been extensively molecularly characterized, is used as a reference cell line in laboratories throughout the world, and is amenable to culture in microtiter plates. A subset of perturbagens were also profiled in the prostate cancer epithelial cell line PC3 and the nonepithelial lines HL60 (leukemia) and SKMEL5 (melanoma). This diversity of cell types provides an opportunity to assess the extent to which results are context dependent.

Concentration and duration of treatment. High-throughput, cell-based small-molecule screens are often performed at a single, relatively high concentration of 10 μM. We adopted this approach as well, given that the optimal concentration is not known for many compounds of potential interest. For some compounds, we used concentrations reported to be effective in cell

[1] Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA. [2] Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [3] Department of Medicine, Children's Hospital Boston, Boston, MA 02115, USA. [4] Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02144, USA. [5] Harvard Medical School, Boston, MA 02115, USA. [6] Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. [7] Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA.

*To whom correspondence should be addressed. E-mail: golub@broad.harvard.edu, justin@broad.mit.edu

†Present address: University of California, San Francisco, CA 94158, USA.

culture or to approximate the maximum attainable plasma concentrations after therapeutic dosing. We also profiled a subset of compounds across a range of concentrations to explore the sensitivity of results to dose.

As with concentration, the duration of compound treatment might also affect the gene-expression profiles. Profiles obtained too early might not yield robust signals—particularly for perturbations that do not directly modulate transcription—and those obtained too late might reflect secondary and tertiary responses. Because our goal was to obtain signatures related to direct mechanisms of action, we selected a relatively early time point (6 hours after compound addition, with a subset also profiled at 12 hours for comparison).

Control perturbations. Every treatment "instance" was defined relative to a control consisting of cells grown in the same plate and treated with vehicle alone. This approach was taken to minimize the impact of batch-to-batch biological and technical variation. Most of the perturbagens were also profiled multiple times.

Overall data. Our data set was thus composed of genomewide mRNA expression data for 164 distinct bioactive small-molecule perturbagens and corresponding vehicle controls applied to human cell lines for short duration. These data were collected in multiple batches over a period of 1 year by means of Affymetrix GeneChip microarrays. A total of 564 gene-expression profiles were produced, representing 453 individual instances (i.e., one treatment and vehicle pair). Full details of the data set are provided as table S1. The data are freely available for download at www.broad.mit.edu/cmap.

Querying the Connectivity Map

The traditional method for identifying small molecules with similar effects on the basis of gene-expression profiles is hierarchical clustering. Indeed, such a strategy was found to be useful for analyzing data from yeast (*7*) and rat tissues (*10*). However, we saw three drawbacks with such an approach. First, with mammalian cell culture, the dominant structure we detected by hierarchical clustering was related to cell type and batch effects (similarity among cells grown at the same time), and this masked the more subtle signals from short-duration treatment with small molecules (fig. S1). Second, a hierarchical clustering approach would require that all profiles be generated on the same microarray platform, limiting future utility. Third, and most important, we required an analytical method that could detect multiple components within the cellular response to a given perturbation.

For these reasons, we adopted a nonparametric, rank-based pattern-matching strategy based on the Kolmogorov-Smirnov statistic (*11*), as we described previously and later formalized in Gene Set Enrichment Analysis (GSEA) (*2*, *12*, *13*). The approach starts with a "query signature" and assesses its similarity to each of the reference expression profiles in the data set. A query signature is any list of genes whose expression is correlated with a biological state of interest. Examples could include genes correlated with a subtype of disease (e.g., drug-resistant versus drug-sensitive leukemia) or regulated by a biological process of interest (e.g., experimental activation of a signaling pathway). Each gene in the query signature carries a sign, indicating whether it is up-regulated or down-regulated. Because the query signature is unitless, it is not tied to any technology platform.

The reference gene-expression profiles in the Connectivity Map data set are also represented in a nonparametric fashion. Each profile is compared to its corresponding intrabatch vehicle-treated control. The genes on the array are rank-ordered according to their differential expression relative to the control; each treatment instance thus gives rise to a rank-ordered list of ~22,000 genes.

The query signature is then compared to each rank-ordered list to determine whether up-regulated query genes tend to appear near the top of the list and down-regulated query genes near the bottom ("positive connectivity") or vice versa ("negative connectivity"), yielding a "connectivity score" ranging from +1 to −1. A null (zero) connectivity score is assigned where the enrichment scores for the up- and down- regulated genes have the same sign. All instances in the database are then ranked according to their connectivity scores; those at the top are most strongly correlated to the query signature, and those at the bottom are most strongly anticorrelated (Fig. 1). (For expression profiles derived from a single technology platform, we obtained similar results using conventional measures of correlation, such as the Pearson correlation coefficient.)

There is no standard approach for estimating the statistical significance of the connections observed. We therefore resorted to the simplest, most empirical, and most transparent test we could devise, and note that the power to detect connections may be greater for compounds with many replicates. Below, we have focused on cases where the precise calculation of *P*-values is not critical to support our conclusions because the observed connections were so striking (and in several cases validated with functional experiments).

Connections Between Small Molecules

HDAC inhibitors. We first determined whether a query signature derived from a class of small molecules could recover those same compounds in the Connectivity Map. A recent report (*14*) described gene-expression responses of T24 (bladder), MDA 435 (breast carcinoma), and MDA 468 (breast carcinoma) cells treated with three histone deacetylase (HDAC) inhibitors: vorinostat (also known as suberoylanilide hydroxamic acid or SAHA), MS-27-275, and trichostatin A. The authors of this study defined a 13-gene signature (8 up-regulated and 5 down-regulated genes; Signature S1) that was used to query our database.

Despite the differences in the cells used to generate the query signature and reference profiles, the two highest-scoring compounds in the Con-

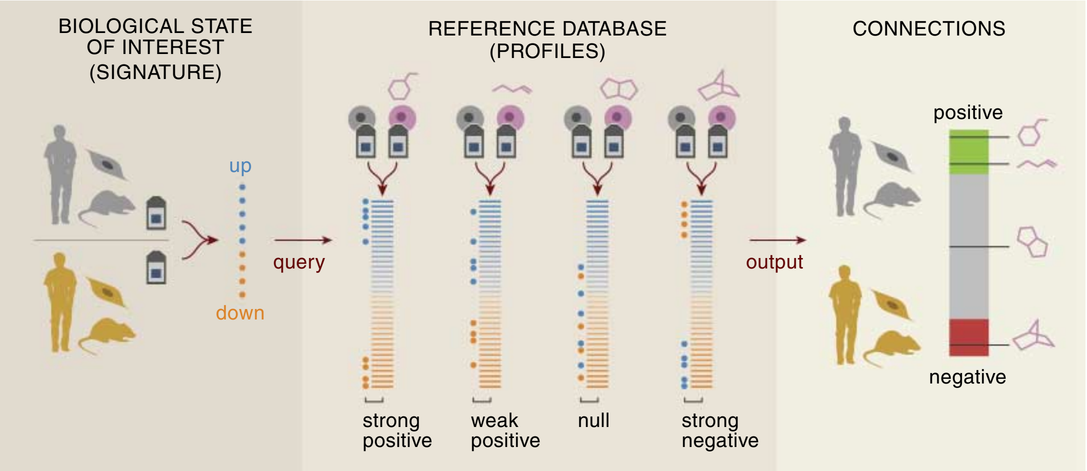


Fig. 1. The Connectivity Map Concept. Gene-expression profiles derived from the treatment of cultured human cells with a large number of perturbagens populate a reference database. Gene-expression signatures represent any induced or organic cell state of interest (**left**). Pattern-matching algorithms score each reference profile for the direction and strength of enrichment with the query signature (**center**). Perturbagens are ranked by this "connectivity score"; those at the top ("positive") and bottom ("negative") are functionally connected with the query state (**right**) through the transitory feature of common gene-expression changes.

nectivity Map were vorinostat and trichostatin A (Fig. 2A). More important, the Connectivity Map also revealed strong connectivity with two structurally distinct compounds, valproic acid (initially developed as an antiseizure drug) and HC toxin, both of which are now known to have HDAC-inhibitory activity but were not used to define the query signature (Fig. 2, A and B). These results indicate that the Connectivity Map would have suggested the HDAC-inhibitory activity of these compounds had it not already been known.

The ability to detect these HDAC inhibitors was not highly sensitive to the precise concentration of drug used to generate the reference profiles. Specifically, the Connectivity Map contains instances of valproic acid at six concentrations (10, 2, and 1 mM; 500, 200, and 50 μM) bracketing the commonly used HDAC-inhibitory level of 1 mM. Only the two lowest concentrations failed to yield a positive connectivity score (Fig. 2A). The results indicate that, at least for this example, connectivity can be established without elaborate optimization of cell type and compound concentration.

Estrogens. We next studied the effects of estrogen, which is known to modulate nuclear hormone signaling. The query signature was taken from a report by an independent group (*15*) in which MCF7 cells were treated with the natural estrogen receptor (ER) ligand, 17β-estradiol (E2). The query signature consisted of 129 genes (40 up- and 89 down-regulated; Signature S2).

The Connectivity Map yielded high positive connectivity scores for all instances of E2 in MCF7 cells. High connectivity scores were also observed for genistein, which is a phytoestrogen (*16*). Weaker connectivity was seen with 17α-estradiol, consistent with its markedly lower affinity for ER than its stereoisomer (*17*) (Fig. 3A).

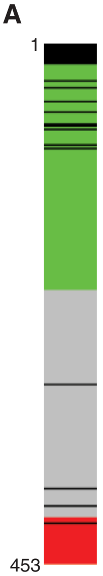
The Connectivity Map also identified compounds with clear negative connectivity, indicating an opposite effect to that of E2. The highest negative connectivity scores came from fulvestrant, a known anti-estrogenic drug (*18*) (Fig. 3B). Tamoxifen and raloxifene, also anti-estrogens, scored negatively, but to a lesser extent (fig. S2). Together, these results indicate that both agonists and antagonists can be discovered directly from the Connectivity Map.

We used estrogen connectivity to explore the impact of physiological context. Such context is known to be particularly important in the study of ER activity, where growing cells in culture medium containing the estromimetic phenol red and supplemented with complete serum (which contains endogenous estrogens) often obscures estrogen stimulation signals as measured with traditional read-outs such as reporter assays or gel shifts. We therefore asked whether the 129-gene estrogen signature might be more robust to the particulars of culture medium composition. Indeed, the presence of phenol red and complete serum had little effect on connectivity scores for E2, even though the query signature was defined under estrogen-free conditions (*15*). Indeed, the connectivity scores were similar to those made in phenol red-free medium with charcoal-stripped serum (ssMCF7; Fig. 3A). However, the anti-estrogen fulvestrant received a null connectivity score in MCF7 cells under estrogen-free conditions, consistent with its "pure antagonist" mode of action (Fig. 3B). Similarly, no robust estrogenic or anti-estrogenic connections were recovered in treatments performed in PC3 or HL60 cells, neither of which expresses ER. These results indicate that although gene-expression signatures can be highly sensitive, some connections will not be found if the reference profiles are collected in cells that lack the appropriate physiological or molecular context.

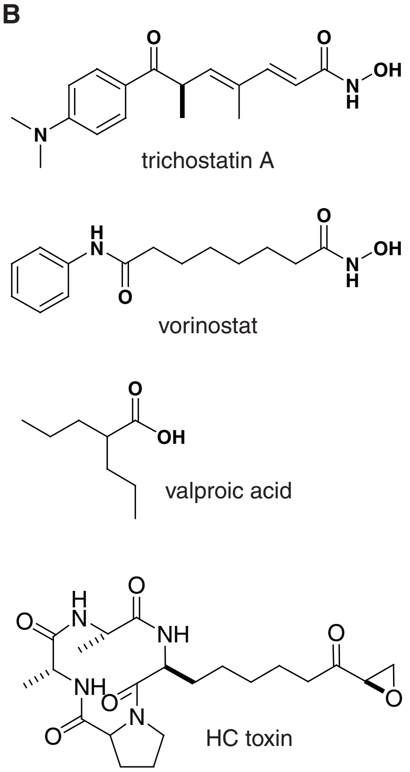
Phenothiazines. We next considered small molecules that do not directly regulate gene expression. We studied the phenothiazine antipyschotics, which as dopamine receptor antagonists and calmodulin inhibitors are not immediately proximal to transcription. Reference profiles were generated for five phenothiazines (chlorpromazine, fluphenazine, prochlorperazine, thioridazine, trifluoperazine) representing three structural subclasses (Fig. 4A). At least three instances of each were produced, mainly in MCF7 cells and at a concentration of 10 μM, with the exception of chlorpromazine, which was profiled three times at 1 μM and only once at 10 μM.

For these experiments, query signatures were generated from a subset of the reference profiles within the Connectivity Map data set itself. A query signature consisting of genes consistently regulated across one instance of each of the five phenothiazines was first established (Signature S3). We then used this signature to assess the recovery of all of the remaining phenothiazine instances from our database. As anticipated, the five instances used to derive the signature received the highest connectivity scores. More important, 10 of the 13 nonsignature instances were also highly ranked (Fig. 4B). The three instances not

Fig. 2. HDAC Inhibitors. (**A**) HDAC inhibitors are highly ranked with an external HDAC inhibitor signature. The "bar-view" is constructed from 453 horizontal lines, each representing an individual treatment instance, ordered by their corresponding connectivity scores with the Glaser *et al.* (*14*) signature (+1, top; –1, bottom). All valproic acid (*n* = 18), trichostatin A (*n* = 12), vorinostat (*n* = 2), and HC toxin (*n* = 1) instances in the data set are colored in black. Colors applied to the remaining instances reflect the sign of their scores (green, positive; gray, null; red, negative). The rank, name [instance id], concentration, cell line, and connectivity score for each of the selected HDAC inhibitor instances is shown. Unabridged results from this query are provided as Result S1. (**B**) Chemical structures.



rank	perturbagen	dose	cell	score
1	vorinostat [1000]	10 μM	MCF7	1
2	trichostatin A [873]	1 μM	MCF7	0.969
3	trichostatin A [992]	100 nM	MCF7	0.931
4	trichostatin A [1050]	100 nM	MCF7	0.929
5	vorinostat [1058]	10 μM	MCF7	0.917
6	trichostatin A [981]	1 μM	MCF7	0.915
7	HC toxin [909]	100 nM	MCF7	0.914
8	trichostatin A [1112]	100 nM	MCF7	0.908
9	trichostatin A [1072]	1 μM	MCF7	0.906
10	trichostatin A [1014]	1 μM	MCF7	0.893
11	trichostatin A [332]	100 nM	MCF7	0.882
12	trichostatin A [331]	100 nM	MCF7	0.846
13	trichostatin A [448]	100 nM	PC3	0.788
14	valproic acid [345]	10 mM	MCF7	0.743
15	valproic acid [23]	1 mM	MCF7	0.735
16	valproic acid [1047]	1 mM	MCF7	0.733
17	trichostatin A [413]	100 nM	ssMCF7	0.725
18	valproic acid [410]	10 mM	HL60	0.725
19	valproic acid [458]	1 mM	PC3	0.680
33	valproic acid [409]	1 mM	HL60	0.634
39	valproic acid [1020]	500 μM	MCF7	0.619
52	valproic acid [346]	2 mM	MCF7	0.582
61	valproic acid [1078]	500 μM	MCF7	0.563
71	valproic acid [629]	1 mM	SKMEL5	0.539
72	valproic acid [347]	500 μM	MCF7	0.539
73	valproic acid [989]	1 mM	MCF7	0.538
76	valproic acid [433]	1 mM	PC3	0.528
89	trichostatin A [364]	100 nM	HL60	0.507
92	valproic acid [497]	1 mM	ssMCF7	0.501
297	valproic acid [348]	50 μM	MCF7	0
388	valproic acid [994]	200 μM	MCF7	0
403	valproic acid [1002]	50 μM	MCF7	0
419	valproic acid [1060]	50 μM	MCF7	-0.537



receiving high connectivity scores were the low-concentration chlorpromazine treatments; these therefore served as useful specificity controls. Similar results were obtained with signatures produced from different phenothiazine instances and with different gene-selection criteria (Signatures S4 to S6; fig. S3). These results show that the common activity of these phenothiazine antipsychotic compounds can be recovered by the Connectivity Map, even when analyzed in nonneural cells. They also demonstrate that the approach is not unduly sensitive to signature-definition parameters.

The phenothiazine query signature did not show strong connectivity with the nonphenothiazine antipsychotics haloperidol and clozapine (Fig. 4C). This is not surprising because, although these antipsychotics ultimately target the same neurotransmitter receptors, the receptors themselves are not expressed in the cell lines used. Indeed, the antipsychotics were included in this data set as an extreme test of the Connectivity Map concept.

The analysis of the phenothiazine query signature did yield consistently strong negative connectivity scores for arachidonic acid (Fig. 4C). Arachidonic acid is the primary substrate for cyclooxygenases and lipoxygenases and is thus a critical precursor for both prostaglandin and leukotreine syntheses. The Connectivity Map result suggests that phenothiazines have an activity that mimics ablation of the arachidonic acid cascade and is therefore entirely consistent with the observation that phenothiazines can inhibit prostaglandin synthesis (*19*). Indeed, more recently, phenothiazine derivatives have been developed as potent dual cyclooxygenase/lipoxygenase inhibitors that exhibit anti-inflammatory activity (*20*). Had this activity of phenothiazines not been previously discovered by serendipity, it would have been systematically revealed by the Connectivity Map.

These findings confirm that even perturbagens not acting immediately proximal to transcription do give rise to distinguishable gene-expression profiles and demonstrate again that the Connectivity Map can reveal complex biological activities. They also show that the Connectivity Map approach can use both internal as well as external query signatures.

Identification of gedunin as an HSP90 inhibitor. We next sought to use the Connectivity Map to generate hypotheses about the mechanism of action of an uncharacterized small molecule. In a separate study, we performed a high-throughput gene expression–based screen for small molecules capable of abrogating the gene-expression signature of androgen receptor (AR) activation in prostate cancer cells. The details of the screen and its biochemical follow-up are described elsewhere (*21*). One of the hits from the screen was the triterpenoid natural product gedunin (*22*) (Fig. 5A), purified from the *Meliacae* family of medicinal plants. The mechanism by which gedunin abrogated AR activity was entirely unknown because this compound has not been extensively characterized.

In an effort to elucidate its mechanism of action, we defined a signature for gedunin (Signature S7) by treating LNCaP prostate cancer cells for 6 hours with the compound, and queried the Connectivity Map. High connectivity scores were found for multiple instances of three heat shock protein 90 (HSP90) inhibitors: geldanamycin, 17-allylamino-geldanamycin, and 17-dimethylamino-geldanamycin (Fig. 5B). As a class, these HSP90 inhibitors showed marked connectivity to the gedunin signature (permutation *P*-value < 0.0001).

This result suggests that gedunin, though structurally dissimilar from known HSP90 inhibitors (Fig. 5A), might impinge upon the HSP90 pathway. Because the stability of AR is known to be dependent upon HSP90 activity, we asked whether AR expression could be diminished by gedunin treatment. Immunoblotting indicated that AR protein, as well as other HSP90-interacting proteins, was nearly entirely eliminated in gedunin-treated LNCaP and Ba/F3 cells (Fig. 5C), consistent with gedunin acting as an inhibitor of HSP90 function. Moreover, mutant interacting proteins such as the BCR-ABL T315I point mutant and the FLT3 internal tandem duplication (ITD) mutant show increased sensitivity to gedunin-mediated inhibition, as is seen upon HSP90 inhibition by geldanamycins (*23*, *24*). Further biochemical studies demonstrated that the mechanism of abrogating HSP90 function was distinct from geldanamycin and its analogs (*21*).

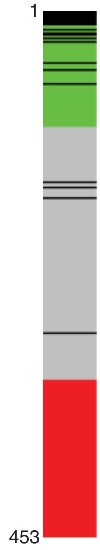
These experiments demonstrate that the Connectivity Map can generate testable hypotheses about the target pathways of poorly characterized small molecules, providing a potentially powerful tool for pharmaceutical development.

Connections with Disease States

We next sought to collect query signatures from disease states and scan the Connectivity Map to identify small molecules that might mimic or suppress that disease.

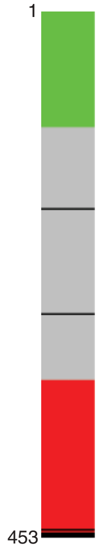
Diet-induced obesity. We made use of a signature for the obese state from a published report (*25*) of the genes differentially expressed in a rat model of diet-induced obesity (Signature S8). The conditions used in that study differed sharply from those used to build the Connectivity Map with respect to RNA source (adipose tissue versus cell lines), treatment duration (65 days versus 6 hours), and species (rat versus human). Despite these differences, instances of three peroxisome proliferator-activated receptor gamma (PPARγ) agonists—the thiazolidindiones (TZD), troglitazone and rosiglitazone, and indometacin (*26*–*28*)—received high connectivity scores (fig. S4). Indeed, all three compounds are potent inducers of adipogenesis in vitro (*26*–*28*). Further, that TZDs promote weight gain in vivo has been widely observed as a consequence of their clinical use

A



rank	perturbagen	dose	cell	score
2	estradiol [988]	100 nM	MCF7	0.936
3	estradiol [373]	10 nM	ssMCF7	0.918
4	genistein [1015]	10 μM	MCF7	0.913
5	estradiol [1079]	10 nM	MCF7	0.899
6	estradiol [1021]	10 nM	MCF7	0.813
8	alpha-estradiol [990]	10 nM	MCF7	0.809
9	alpha-estradiol [403]	10 nM	ssMCF7	0.807
10	estradiol [414]	10 nM	ssMCF7	0.794
11	estradiol [121]	10 nM	MCF7	0.758
12	genistein [1073]	10 μM	MCF7	0.753
13	genistein [638]	10 μM	MCF7	0.730
17	alpha-estradiol [1048]	10 nM	MCF7	0.646
20	genistein [268]	1 μM	MCF7	0.619
21	estradiol [365]	100 nM	MCF7	0.610
25	genistein [382]	10 μM	MCF7	0.561
27	genistein [267]	1 μM	MCF7	0.552
46	alpha-estradiol [122]	10 nM	MCF7	0.435
51	estradiol [387]	10 nM	HL60	0.421
64	estradiol [782]	10 nM	HL60	0.376
148	alpha-estradiol [702]	10 nM	PC3	0
152	genistein [703]	10 μM	PC3	0
162	alpha-estradiol [762]	10 nM	MCF7	0
278	estradiol [665]	10 nM	PC3	0

B



rank	perturbagen	dose	cell	score
171	fulvestrant [704]	1 μM	PC3	0
261	fulvestrant [523]	1 μM	ssMCF7	0
447	fulvestrant [367]	1 μM	MCF7	-0.749
450	fulvestrant [310]	10 nM	MCF7	-0.843
451	fulvestrant [985]	1 μM	MCF7	-0.961
452	fulvestrant [1076]	10 nM	MCF7	-0.989
453	fulvestrant [1043]	1 μM	MCF7	-1

Fig. 3. Estrogen receptor agonists and antagonists. (**A**) Estrogen receptor agonists are highly ranked with an external 17β-estradiol signature. Barview (as in Fig. 2) with all estradiol (*n* = 10), alpha-estradiol (*n* = 7), and genistein (*n* = 7) instances shown. (**B**) A pure antiestrogen is negatively connected. Barview with all fulvestrant (*n* = 7) instances shown. Unabridged results from this query are provided as Result S2.

as oral antidiabetic agents and is considered a major drawback of their use (*29*). The Connectivity Map would have predicted this particular adverse effect.

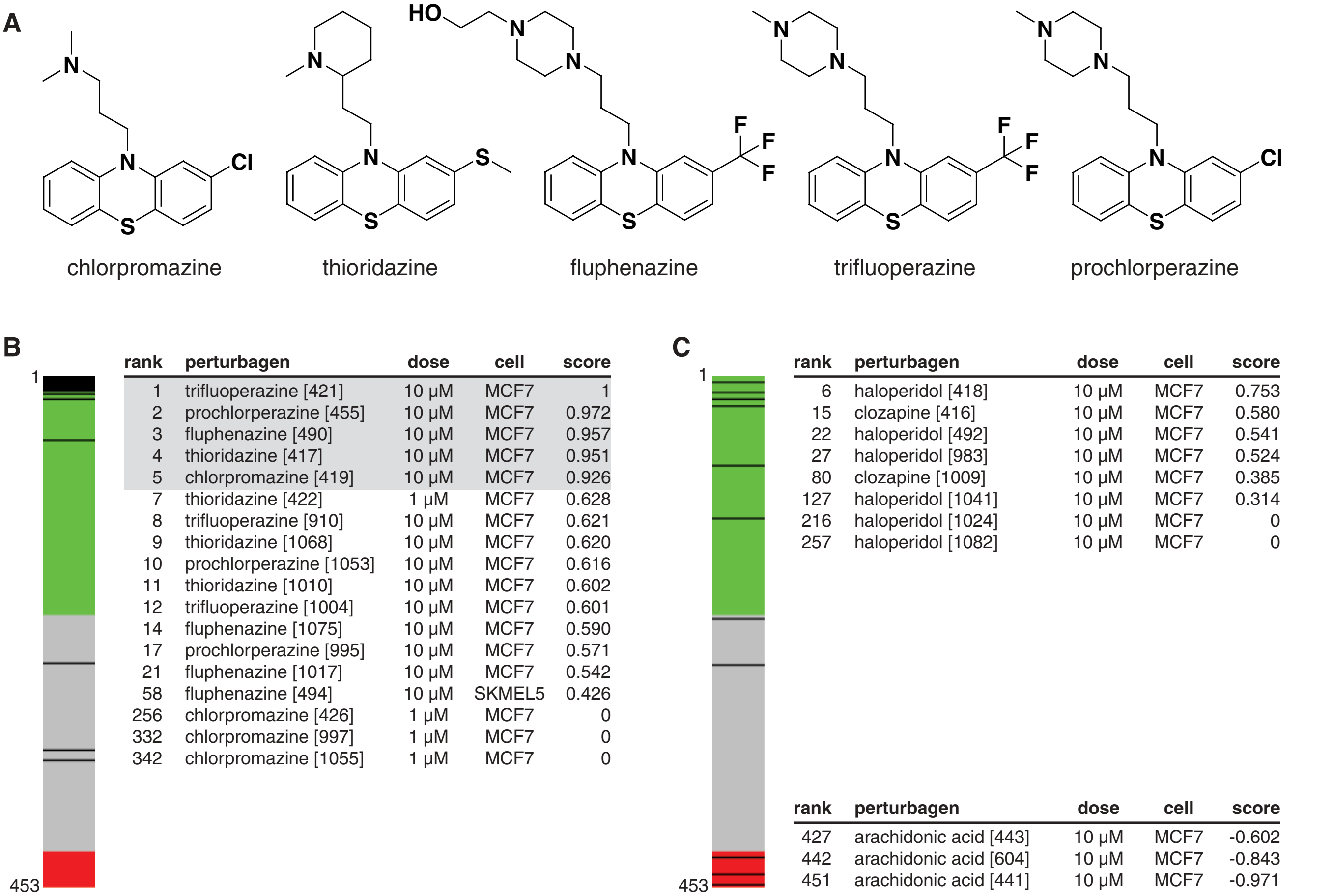
These results must be tempered, however, because they derive solely from PC3—the only cell line in our panel to express PPARγ at high levels (*30*)—and TZD and indometacin instances made in all other cell lines yielded null or negative scores. Clearly, these connections would not have been made had this particular cellular context not been represented. Of note, the other known PPARγ agonist in our collection, 15-delta prostaglandin J2 (*26*), received a null score even in PC3, although the entire set of agonists (including 15-delta prostaglandin J2) still showed significant enrichment as a class (permutation *P*-value = 0.0021). Overall, it is notable that a signature derived from rat adipose tissue after many weeks of treatment can generate connections with small molecules applied acutely to epithelial cells in culture.

Alzheimer's disease. We next explored query signatures for Alzheimer's disease (AD). AD is the most common cause of dementia in the elderly, but its pathogenesis is poorly understood and effective therapies remain elusive. We made use of two independent reports of the gene-expression changes in brain tissue from AD patients.

The first signature consisted of 40 genes identified through a comparison of hippocampus from AD and normal brain (*31*) (Signature S9). The second, derived from the comparison between cerebral cortex from AD brain and age-matched controls, contained 25 genes (*32*) (Signature S10). Although there were no genes in common between these two query signatures, both yielded statistically significant negative connectivity with the two independent instances of 4,5-dianilinophthalimide (DAPH) in the Connectivity Map (fig. S5). No other compound in the database shared this behavior.

DAPH was recently identified in a cell-free screen for small molecules that could reverse the formation of fibrils (specifically, decreasing the β-sheet content of aggregating Aβ1-42 peptide) thought to be responsible for the accelerated neuronal cell death in the brains of AD patients (*33*). Indeed, a variety of new DAPH analogs have since been synthesized as potential treatments for AD (*34*). Our observations strengthen the candidacy of DAPH as a potential AD therapeutic and further illustrate the potential of the Connectivity Map to generate novel, unbiased hypotheses concerning the pharmacologic modulation of disease states.

Dexamethasone resistance in ALL. As a final example, we considered one of the most vexing problems in cancer chemotherapy: drug resistance. Specifically, we explored resistance to the glucocorticoid dexamethasone in children with acute lymphoblastic leukemia (ALL). Dexamethasone resistance has been observed both in vivo and in primary leukemia cells grown in short-term culture (*35*). We defined a gene-expression signature of dexamethasone sensitivity (Signature S11) by comparing bone-marrow leukemic cells from patients exhibiting either dexamethasone sensitivity or resistance in vitro. The details of this signature are reported elsewhere (*36*).



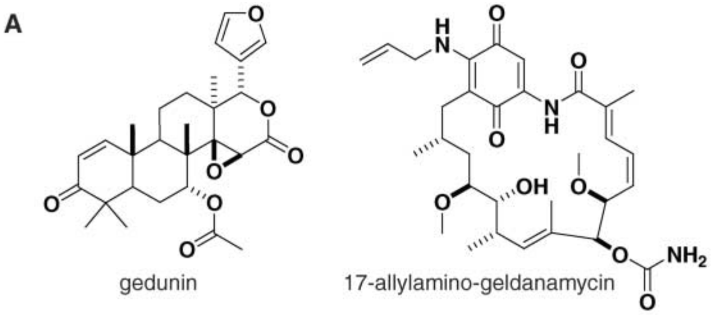
rank	perturbagen	dose	cell	score
1	trifluoperazine [421]	10 µM	MCF7	1
2	prochlorperazine [455]	10 µM	MCF7	0.972
3	fluphenazine [490]	10 µM	MCF7	0.957
4	thioridazine [417]	10 µM	MCF7	0.951
5	chlorpromazine [419]	10 µM	MCF7	0.926
7	thioridazine [422]	1 µM	MCF7	0.628
8	trifluoperazine [910]	10 µM	MCF7	0.621
9	thioridazine [1068]	10 µM	MCF7	0.620
10	prochlorperazine [1053]	10 µM	MCF7	0.616
11	thioridazine [1010]	10 µM	MCF7	0.602
12	trifluoperazine [1004]	10 µM	MCF7	0.601
14	fluphenazine [1075]	10 µM	MCF7	0.590
17	prochlorperazine [995]	10 µM	MCF7	0.571
21	fluphenazine [1017]	10 µM	MCF7	0.542
58	fluphenazine [494]	10 µM	SKMEL5	0.426
256	chlorpromazine [426]	1 µM	MCF7	0
332	chlorpromazine [997]	1 µM	MCF7	0
342	chlorpromazine [1055]	1 µM	MCF7	0

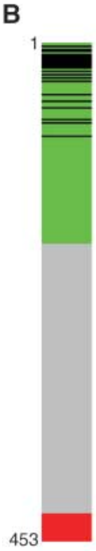
rank	perturbagen	dose	cell	score
6	haloperidol [418]	10 µM	MCF7	0.753
15	clozapine [416]	10 µM	MCF7	0.580
22	haloperidol [492]	10 µM	MCF7	0.541
27	haloperidol [983]	10 µM	MCF7	0.524
80	clozapine [1009]	10 µM	MCF7	0.385
127	haloperidol [1041]	10 µM	MCF7	0.314
216	haloperidol [1024]	10 µM	MCF7	0
257	haloperidol [1082]	10 µM	MCF7	0

rank	perturbagen	dose	cell	score
427	arachidonic acid [443]	10 µM	MCF7	-0.602
442	arachidonic acid [604]	10 µM	MCF7	-0.843
451	arachidonic acid [441]	10 µM	MCF7	-0.971

Fig. 4. Phenothiazine connections. (**A**) Chemical structures. Three structural subclasses are shown: with a piperazine group in the side chain (fluphenazine, trifluoperazine, prochlorperazine), with a piperidine ring in the side chain (thioridazine), and with an aliphatic side chain (chlorpromazine). (**B**) Recovery of phenothiazine instances with an internal phenothazine signature. Barview (as in Fig. 2) with all thioridazine (*n* = 4), chlorpromazine (*n* = 4), fluphenazine (*n* = 4), trifluoperazine (*n* = 3), and prochlorperazine (*n*=3) instances shown. The instances used to generate the signature are shaded. (**C**) Ranking of nonphenothiazine antipsychotics and arachidonic acid instances with the phenothiazine signature. Barview showing all haloperidol (*n* = 6), clozapine (*n* = 2), and arachidonic acid (*n* = 3) instances. Permutation *P*-values are 0.1428, 0.0621, and 0.0002, respectively. Unabridged results from this query are provided as Result S3.

When the signature of dexamethasone sensitivity was used to query the Connectivity Map, we found strong connectivity to the mTOR inhibitor sirolimus (also known as rapamycin) (Fig. 6A). This result suggested that sirolimus might revert dexamethasone resistance. Indeed, treatment of the lymphoid cell line CEM-c1 with sirolimus conferred dexamethasone sensitivity to this otherwise resistant cell line, reducing the median inhibitory concentration (IC_{50}) by a factor of more than 50 (Fig. 6B). Additional experiments indicated that this activity was mTOR dependent, resulting in apoptosis mediated through downregulation of the anti-apoptotic protein MCL1 (*36*). Whatever the mechanism, the result from





rank	perturbagen	dose	cell	score
5	17-allylamino-geldanamycin [1064]	1 μM	MCF7	0.893
8	17-allylamino-geldanamycin [947]	1 μM	MCF7	0.866
9	17-allylamino-geldanamycin [1063]	1 μM	MCF7	0.858
12	geldanamycin [972]	1 μM	MCF7	0.837
13	17-allylamino-geldanamycin [1056]	1 μM	MCF7	0.836
14	17-allylamino-geldanamycin [831]	1 μM	MCF7	0.830
15	17-allylamino-geldanamycin [1044]	1 μM	MCF7	0.822
17	17-dimethylamino-geldanamycin [1051]	100 nM	MCF7	0.808
18	geldanamycin [1008]	1 μM	MCF7	0.787
19	geldanamycin [1066]	1 μM	MCF7	0.784
20	17-allylamino-geldanamycin [916]	1 μM	MCF7	0.769
21	17-allylamino-geldanamycin [998]	1 μM	MCF7	0.767
23	17-allylamino-geldanamycin [986]	1 μM	MCF7	0.748
25	17-allylamino-geldanamycin [1006]	1 μM	MCF7	0.725
27	geldanamycin [611]	1 μM	MCF7	0.710
28	17-allylamino-geldanamycin [1005]	1 μM	MCF7	0.699
30	geldanamycin [864]	1 μM	MCF7	0.686
33	17-allylamino-geldanamycin [428]	1 μM	MCF7	0.654
34	17-dimethylamino-geldanamycin [993]	100 nM	MCF7	0.651
37	geldanamycin [593]	1 μM	MCF7	0.639
48	17-allylamino-geldanamycin [521]	1 μM	ssMCF7	0.571
54	17-allylamino-geldanamycin [381]	1 μM	MCF7	0.553
60	17-allylamino-geldanamycin [221]	100 nM	MCF7	0.540
71	17-allylamino-geldanamycin [505]	1 μM	SKMEL5	0.517
74	17-allylamino-geldanamycin [432]	1 μM	PC3	0.504
87	17-allylamino-geldanamycin [450]	1 μM	PC3	0.484

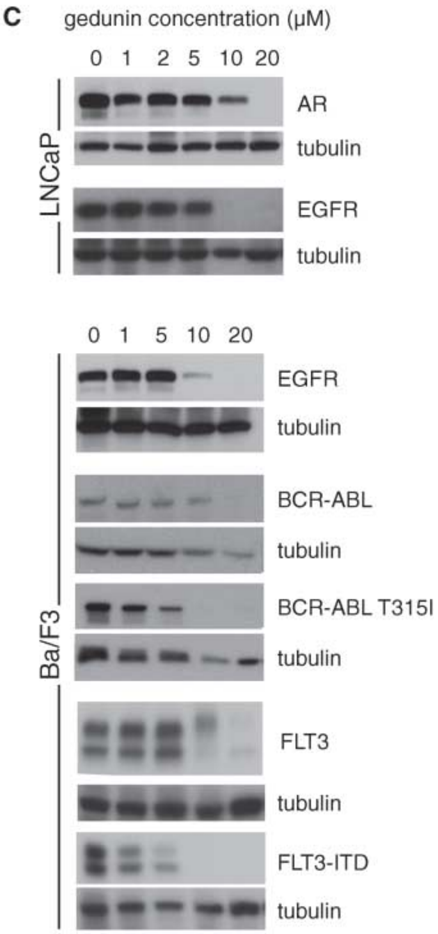
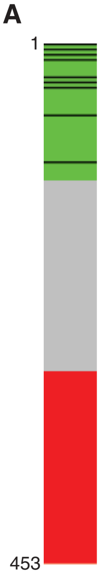


Fig. 5. Gedunin modulates the HSP90 pathway. (**A**) Chemical structure of gedunin and 17-allylamino-geldanamycin. (**B**) Gedunin is connected with geldanamycin and its analogs. Barview (as in Fig. 2) showing all 17-allylamino-geldanamycin (n = 18), geldanamycin (n = 6), and 17-dimethylamino-geldanamycin (n = 2) instances for the gedunin signature. Unabridged results from this query are provided as Result S7. (**C**) Gedunin lowers the levels of HSP90-interacting proteins, including the androgen receptor (AR), in LNCaP cells and Ba/F3 cells ectopically expressing them. Mutant HSP90-interacting proteins (BCR-ABL T315I point mutant and the FLT3-ITD internal tandem duplication mutant) show increased sensitivity to gedunin treatment. EGFR, epidermal growth factor receptor.



rank	perturbagen	dose	cell	score
1	sirolimus [1059]	100 nM	MCF7	1
2	sirolimus [987]	100 nM	MCF7	0.979
7	sirolimus [1045]	100 nM	MCF7	0.930
11	sirolimus [1022]	100 nM	MCF7	0.897
16	sirolimus [1080]	100 nM	MCF7	0.880
30	sirolimus [1001]	100 nM	MCF7	0.810
35	sirolimus [402]	100 nM	ssMCF7	0.793
39	sirolimus [362]	100 nM	HL60	0.778
64	sirolimus [921]	100 nM	MCF7	0.687
104	sirolimus [326]	100 nM	MCF7	0.530

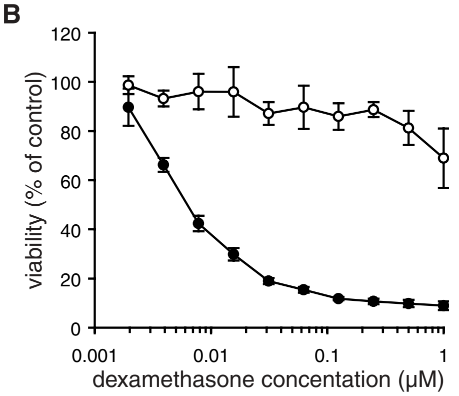


Fig. 6. Sirolimus reverses glucocorticoid resistance in acute lymphoblastic leukemia. (**A**) Barview (as in Fig. 2) showing all 10 sirolimus instances. Permutation P-value for this set of instances is <0.0001. Unabridged results from this query are provided as Result S11. (**B**) The effect of a combination of sirolimus and dexamethasone on the viability of glucocorticoid-resistant lymphoid cells. CEM-c1 cells were treated with 10 nM sirolimus and various concentrations of dexamethasone (closed circles) or dexamethasone alone (open circles) for 72 hours. Cell viability was assessed by MTT reduction and expressed relative to untreated control cultures. Plot shows means and standard deviations for triplicate determinations from one representative experiment.

the Connectivity Map immediately suggests that sirolimus should be tested in a clinical trial of ALL patients with dexamethasone resistance. Sirolimus is already FDA approved as an immunosuppressant and is well tolerated in children, and the clinical prognosis of dexamethasone-resistant ALL is poor (*37*–*40*). This example demonstrates that the Connectivity Map is one approach to the rapid identification of new potential uses for existing drugs.

Discussion

The value of a Connectivity Map depends on many open questions. How many distinct cellular pathways and states actually exist? How many cell types must be studied to provide sufficient diversity? How many perturbagens (small molecules, inhibitory RNAs, open reading frames) would need to be characterized to provide substantial coverage? How many concentrations, time points, and replicates would be required to provide reliable data? What analytical tools will be needed to interpret the data and determine precise estimates of statistical significance and false-positive rates? And, most important, what will be the biomedical value of the data? Only empirical evidence will resolve these issues.

Although only a first step, our results are encouraging. They show that genomic signatures can be used to recognize drugs with common mechanisms of action (HDAC inhibitors and estrogen receptor modulators), discover unknown mechanisms of actions (gedunin as an HSP90 inhibitor), and identify potential new therapeutics (the ability of sirolimus to overcome dexamethasone resistance in ALL). Our findings also reveal that signatures are often conserved across diverse cell types and settings (the signature of dexamethasone resistance was defined in bone-marrow samples but searched against profiles from the MCF7 breast cancer line). At the same time, the results demonstrate the limitations of using only a few cell lines (the signature of estradiol was not detected in cells that lack estrogen receptors) or only a few concentrations (chlorpromazine was not recognized as a phenothiazine at 1 μM). It is also likely that our methodologies can still be refined. Indeed, alternative signature-based pattern-matching methods have been developed [e.g., (*41*)]. In addition, the interpretation of results depends on the ability to confidently call connections. More rigorous methods for the estimation of statistical significance are therefore probably also required, especially as the size of the database grows. But overall, the basic features of our approach appear to work well. We have, therefore, created a Web-based tool (www.broad.mit.edu/cmap) to allow researchers to perform their own Connectivity Map analyses with user-defined signatures in real time.

On the basis of the results of this pilot study, we propose that a sensible next step would be the generation of an expanded Connectivity Map as a community resource project in the spirit of other genomic efforts. An initial goal might be to profile all FDA-approved drugs and inhibitory RNAs targeting a large collection of genes in perhaps 10 diverse cell lines. Further goals would depend on the utility of the data. Ultimately, it will be interesting to explore whether it is possible to create a truly comprehensive catalog that begins to saturate all possible cellular states. In the meanwhile, even an incomplete Connectivity Map will likely accelerate progress in characterizing new chemical entities, finding new uses for existing drugs, and understanding the molecular mechanisms of disease.

References and Notes

1. J. L. DeRisi, V. R. Iyer, P. O. Brown, *Science* **278**, 680 (1997).
2. J. Lamb *et al.*, *Cell* **114**, 323 (2003).
3. T. R. Golub *et al.*, *Science* **286**, 531 (1999).
4. C. M. Perou *et al.*, *Nature* **406**, 747 (2000).
5. S. L. Pomeroy *et al.*, *Nature* **415**, 436 (2002).
6. L. J. van 't Veer *et al.*, *Nature* **415**, 530 (2002).
7. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).
8. B. Ganter *et al.*, *J. Biotechnol.* **119**, 219 (2005).
9. M. R. Fielden *et al.*, *Toxicol. Pathol.* **33**, 675 (2005).
10. J. F. Waring *et al.*, *Toxicol. Appl. Pharmacol.* **175**, 28 (2001).
11. M. Hollander, D. Wolfe, *Nonparametric Statistical Methods* (Wiley, New York, ed. 2, 1999), pp. 178–185.
12. V. K. Mootha *et al.*, *Nat. Genet.* **34**, 267 (2003).
13. A. Subramanian *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545 (2005).
14. K. B. Glaser *et al.*, *Mol. Cancer Ther.* **2**, 151 (2003).
15. J. Frasor *et al.*, *Cancer Res.* **64**, 1522 (2004).
16. P. M. Martin, K. B. Horwitz, D. S. Ryan, W. L. McGuire, *Endocrinology* **103**, 1860 (1978).
17. D. P. Edwards, W. L. McGuire, *Endocrinology* **107**, 884 (1980).
18. A. E. Wakeling, M. Dukes, J. Bowler, *Cancer Res.* **51**, 3867 (1991).
19. J. Y. Vanderhoek, M. B. Feinstein, *Mol. Pharmacol.* **16**, 171 (1979).
20. B. L. Mylari, T. J. Carty, P. F. Moore, W. J. Zembrowski, *J. Med. Chem.* **33**, 2019 (1990).
21. H. Hieronymus *et al.*, *Cancer Cell*, in press.
22. S. A. Khalid, H. Duddeck, M. Gonzalez-Sierra, *J. Nat. Prod.* **52**, 922 (1989).
23. M. E. Gorre, K. Ellwood-Yen, G. Chiosis, N. Rosen, C. L. Sawyers, *Blood* **100**, 3041 (2002).
24. Q. Yao *et al.*, *Clin. Cancer Res.* **9**, 4483 (2003).
25. I. P. Lopez *et al.*, *Obes. Res.* **11**, 188 (2003).
26. B. M. Forman *et al.*, *Cell* **83**, 803 (1995).
27. J. M. Lehmann *et al.*, *J. Biol. Chem.* **270**, 12953 (1995).
28. J. M. Lehmann, J. M. Lenhard, B. B. Oliver, G. M. Ringold, S. A. Kliewer, *J. Biol. Chem.* **272**, 3406 (1997).
29. T. M. Larsen, S. Toubro, A. Astrup, *Int. J. Obes.* **27**, 147 (2003).
30. T. Kubota *et al.*, *Cancer Res.* **58**, 3344 (1998).
31. R. Hata *et al.*, *Biochem. Biophys. Res. Commun.* **284**, 310 (2001).
32. R. Ricciarelli *et al.*, *IUBMB Life* **56**, 349 (2004).
33. B. J. Blanchard *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14326 (2004).
34. E. J. Hennessy, S. L. Buchwald, *J. Org. Chem.* **70**, 7371 (2005).
35. W. J. Tissing, J. P. Meijerink, M. L. den Boer, R. Pieters, *Leukemia* **17**, 17 (2003).
36. G. Wei *et al.*, *Cancer Cell*, in press.
37. R. Pieters *et al.*, *Lancet* **338**, 399 (1991).
38. T. Hongo, S. Yajima, M. Sakurai, Y. Horikoshi, R. Hanada, *Blood* **89**, 2959 (1997).
39. G. J. Kaspers *et al.*, *Blood* **90**, 2723 (1997).
40. G. J. Kaspers *et al.*, *Blood* **92**, 259 (1998).
41. G. Natsoulis *et al.*, *Genome Res.* **15**, 724 (2005).

42. We thank S. Schreiber, E. Scolnick, D. Altshuler, B. Wagner, B. Ebert, N. Tolliday, M. Brown, B. Wong, and members of the Broad Cancer and Chemical Biology Programs. This work was supported in part by grants from the National Cancer Institute, Howard Hughes Medical Institute, and The Paul G. Allen Family Foundation.

Supporting Online Material

www.sciencemag.org/cgi/content/full/313/5795/1929/DC1
Materials and Methods
Figs. S1 to S5
Table S1
Signatures S1 to S11
Results S1 to S11
References

24 July 2006; accepted 6 September 2006
10.1126/science.1132939

Structure of the 70*S* Ribosome Complexed with mRNA and tRNA

Maria Selmer,* Christine M. Dunham,* Frank V. Murphy IV, Albert Weixlbaumer, Sabine Petry, Ann C. Kelley, John R. Weir, V. Ramakrishnan†

The crystal structure of the bacterial 70*S* ribosome refined to 2.8 angstrom resolution reveals atomic details of its interactions with messenger RNA (mRNA) and transfer RNA (tRNA). A metal ion stabilizes a kink in the mRNA that demarcates the boundary between A and P sites, which is potentially important to prevent slippage of mRNA. Metal ions also stabilize the intersubunit interface. The interactions of E-site tRNA with the 50*S* subunit have both similarities and differences compared to those in the archaeal ribosome. The structure also rationalizes much biochemical and genetic data on translation.

A major breakthrough for our mechanistic understanding of translation was achieved some years ago when high-resolution structures of the 50*S* and 30*S* ribosomal subunits were solved (*1*, *2*). Progress has also been made in obtaining structural data on the whole ribosome. The subunit structures were used to facilitate interpretation of maps at 5.5 Å resolution of the whole 70*S* ribosome complexed with mRNA and tRNA (*3*). More recently, the structure of the *Escherichia coli* ribosome was solved at 3.5 Å resolution (*4*). At the same

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: ramak@mrc-lmb.cam.ac.uk