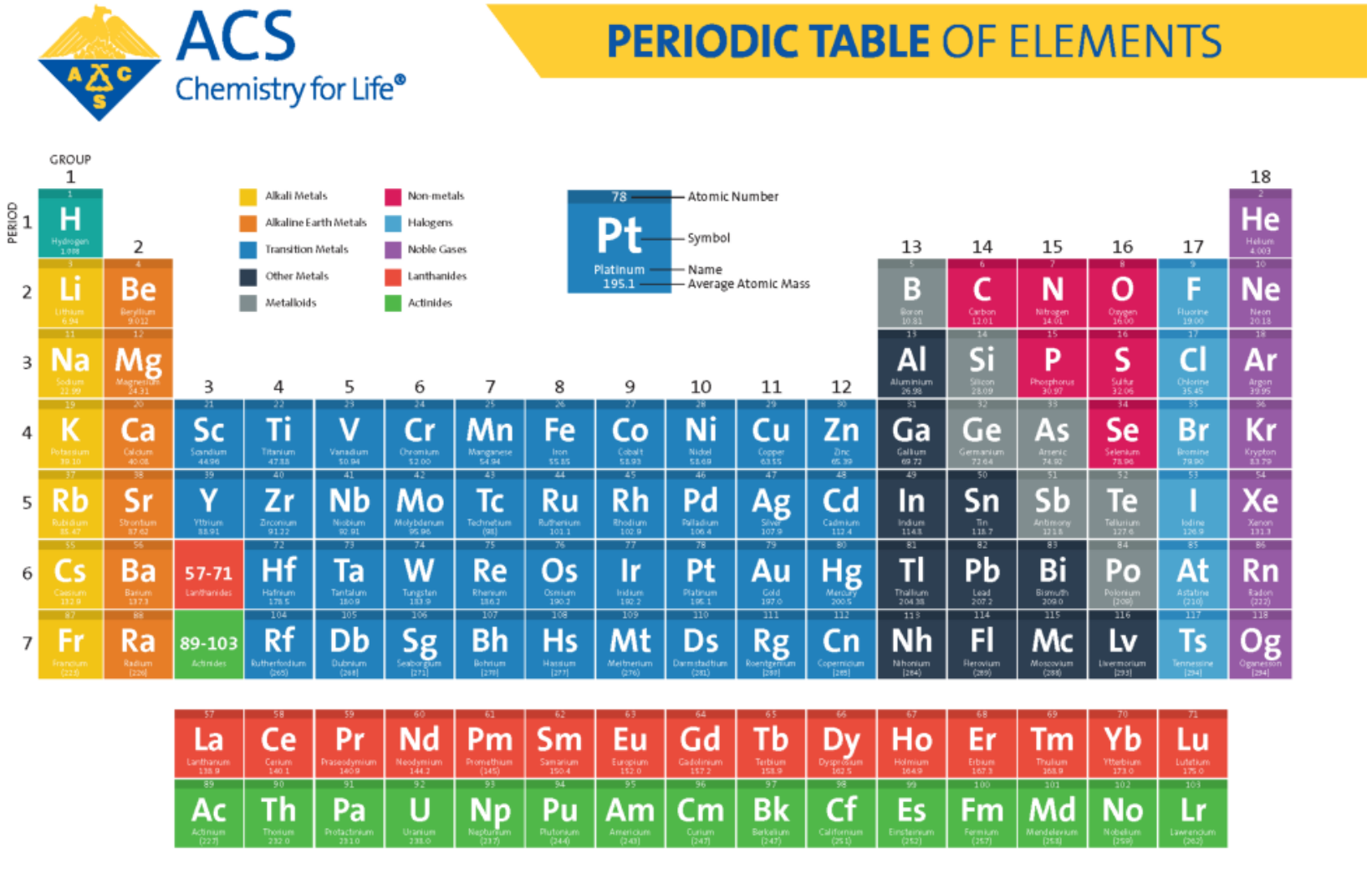
# Connectivity Map (CMAP): Library of Perturbagen-derived Gene Signatures

Aik Choon Tan

10/30/2025

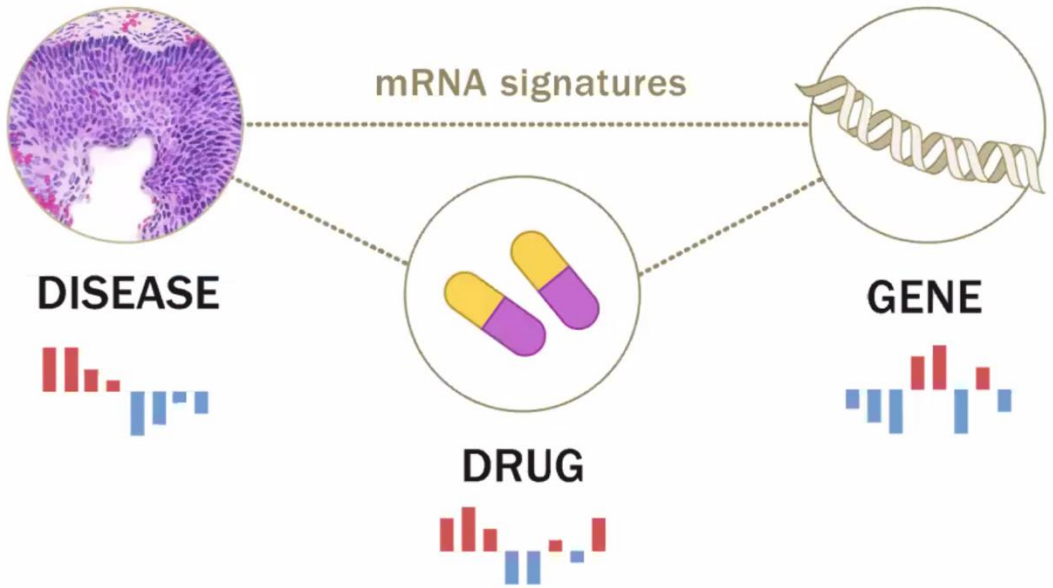
# Periodic Table of Elements (Look-up Table)



# Outline

- CMap Concept
- Pattern-matching algorithm
- Next-generation CMap
- Applications
- Beyond Gene Expression-based CMap

# Connectivity Map: a library of perturbational responses



# Library of perturbational signatures

## Desired attributes of library

- Comprehensive (like library of Congress)
- Information-rich readouts
  (not optimized to particular questions)
- Easy to look things up (like Google search)
- Easy to compare to non-CMap data
  (like image search)
- Accessible to bench experimentalists
  and computationalists

Connectivity Map

# 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*]

(Science 2006)

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.

# Connectivity Map Concept

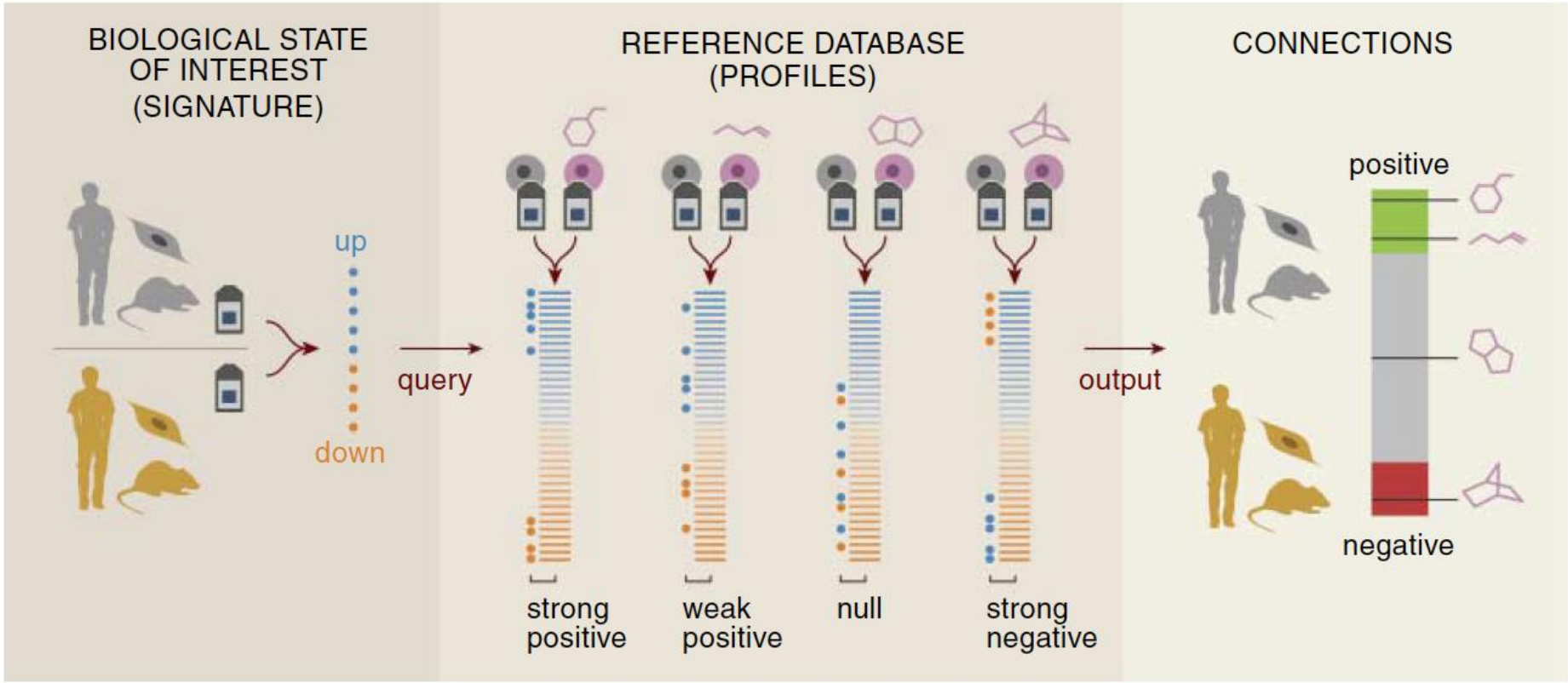
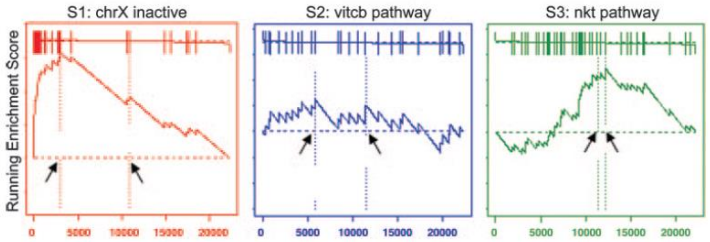



**Fig. 1.** The Connectivity Map Concept. Gene-expression profiles derived from the treatment of cultured human cells with a large number of perturbagens populate a reference database. Gene-expression signatures represent any induced or organic cell state of interest (**left**). Pattern-matching algorithms 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.

# Pattern-Matching Algorithm

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.




***Connectivity Analytics***

Treatment instances were rank ordered with respect to a given query signature using a gene-set enrichment metric based on the Kolmogorov-Smirnov statistic (*1, 2*), as follows. For each instance *i*, compute an enrichment score for the set of probe sets ('tags') representing the up- or down- regulated genes in the signature; $ks^i_{up}$ and $ks^i_{down}$, respectively. Let *n* be the total number of probe sets (22,283) and *t* be the number of tags. Construct a vector *V* of the position (1… *n*) of each tag in the ordered list of all probe sets (see Data Preprocessing, above) and sort these components in ascending order such that *V(j)* is the position of tag *j*, where *j* = 1, 2, …, *t*. Compute the following two values:

$$a = \max_{j=1}^{t}\left[\frac{j}{t} - \frac{V(j)}{n}\right]$$

$$b = \max_{j=1}^{t}\left[\frac{V(j)}{n} - \frac{(j-1)}{t}\right]$$

… and set $ks^i = a$, if $a > b$ or $ks^i = -b$ if $b > a$. The connectivity score $S^i$ is set to zero where $ks^i_{up}$ and $ks^i_{down}$ have the same algebraic sign. Otherwise, set $s^i = ks^i_{up} - ks^i_{down}$, $p = \max(s^i)$ and $q = \min(s^i)$ across all instances. The connectivity score $S^i$ for the non-zero instances is defined as $s^i / p$ where $s^i > 0$, or $-(s^i / q)$ where $s^i < 0$. Instances are then ranked in descending order of *S* and $ks_{up}$. Finding multiple independent instances of the same perturbagen with high (or low) rankings was taken to indicate positive (or negative) connectivity between that perturbagen and the biology represented in the query signature.

The significance of a particular distribution of a set of instances in the ordered list of all instances was estimated by permutation, as follows. The Kolmogorov-Smirnov statistic is computed for the set of *t* instances of interest in the ordered list of *n* instances as above, giving an enrichment score $KS_0$. Then, for each of *r* trials, select *t* instances at random from the set of *n* instances and compute $KS_m$, and count the number of times *s* that $|KS_m| \geq |KS_0|$ is true. The frequency of this event (*s* / *r*) can be taken as a (two-sided) p-value. We set *r*=10,000.

# Some other Pattern-Matching Algorithms

**TABLE 1**
**The Similarities and Differences of Pattern-Matching Algorithms Used in Connectivity Map Analysis**

| Pattern-Matching Algorithms | Query Signature | Descriptions | Technical Limitations | References |
|---|---|---|---|---|
| KS test | • Upregulated gene symbols<br>• Downregulated gene symbols | • Nonparametric<br>• Genes are converted to rank | • Query gene signatures must be rank-ordered<br>• The CS is set to zero if ES for the upregulated and downregulated genes have the same algebraic sign | Lamb et al [1], 2006 |
| WTCS | • Upregulated gene symbols<br>• Downregulated gene symbols | • Nonparametric<br>• Genes are converted to rank | The CS is set to zero if ES for the upregulated and downregulated genes have the same algebraic sign | Subramanian et al [2], 2017 |
| ssCMap ordered | • Upregulated gene symbols<br>• Downregulated gene symbols<br>• Fold change values | Genes are ordered using the absolute value of their fold change values | Fold change value of the query genes are required | Zhang and Gant [7], 2008 |
| ssCMap unordered | • Upregulated gene symbols<br>• Downregulated gene symbols | No particular ordering of the genes; +1 for upregulated genes and −1 for downregulated genes | The significance of which genes are more important could not be determined because all genes take on the value of either +1 (upregulated) or −1 (downregulated) | Zhang and Gant [7], 2008 |
| XSum | • Upregulated gene symbols<br>• Downregulated gene symbols | Limited to the top *N* and bottom *N* of the genes in reference profile | Execution of XSum will collapse if there are no overlapping query gene signatures with the genes in the reference profile | Cheng et al [8], 2014 |
| XCos | • Upregulated gene symbols<br>• Downregulated gene symbols<br>• Fold change values | Limited to the top *N* and bottom *N* of the genes in reference profile | • Fold change value of the query genes is a prerequisite for XCos execution<br>• Execution of XCos will collapse if there are nonoverlapping query gene signatures with the genes in the reference profile<br>• Must have at least 2 query genes overlapping with the genes in the reference profile | Cheng et al [9], 2013 |

(Lee, Cheong, Tan, *Adv. Mol. Path.* 2019)

# CMap Results

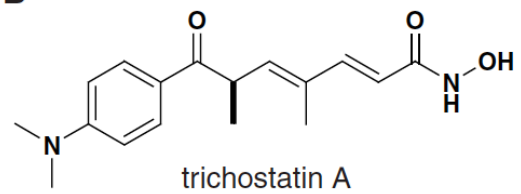
*HDAC inhibitors*. We first determined whether a query signature derived from a class of small molecules could recover those same compounds in the Connectivity Map. A recent report (14) described gene-expression responses of T24 (bladder), MDA 435 (breast carcinoma), and MDA 468 (breast carcinoma) cells 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 downregulated genes; Signature S1) that was used to query our database.

**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.

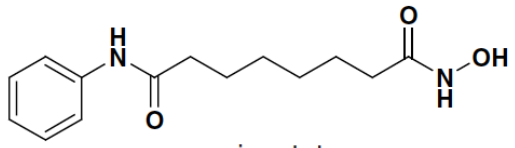
**A**


| 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 |

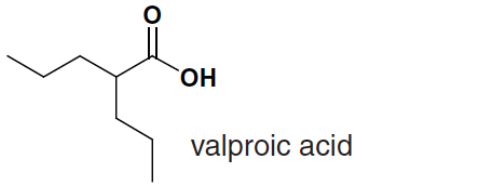
**B**



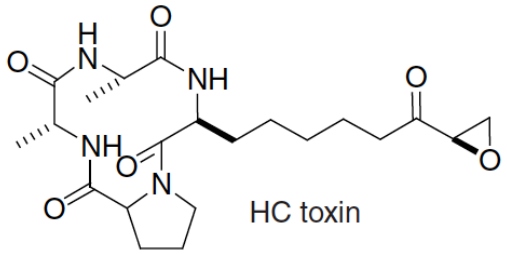

trichostatin A




vorinostat

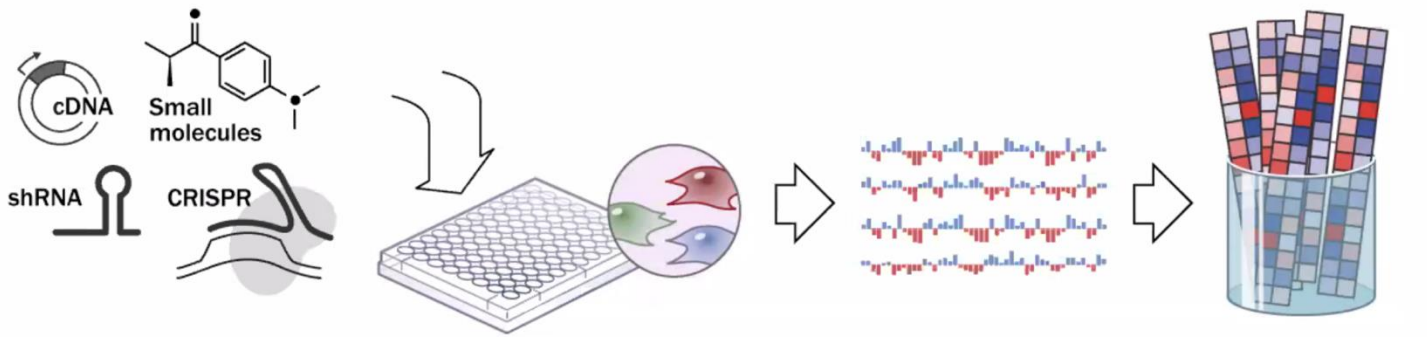



valproic acid




HC toxin

# Connectivity Map




**Pilot: 154 drugs**

NIH LINCS PROGRAM

**> 3,000,000 profiles**
**all publicly available**

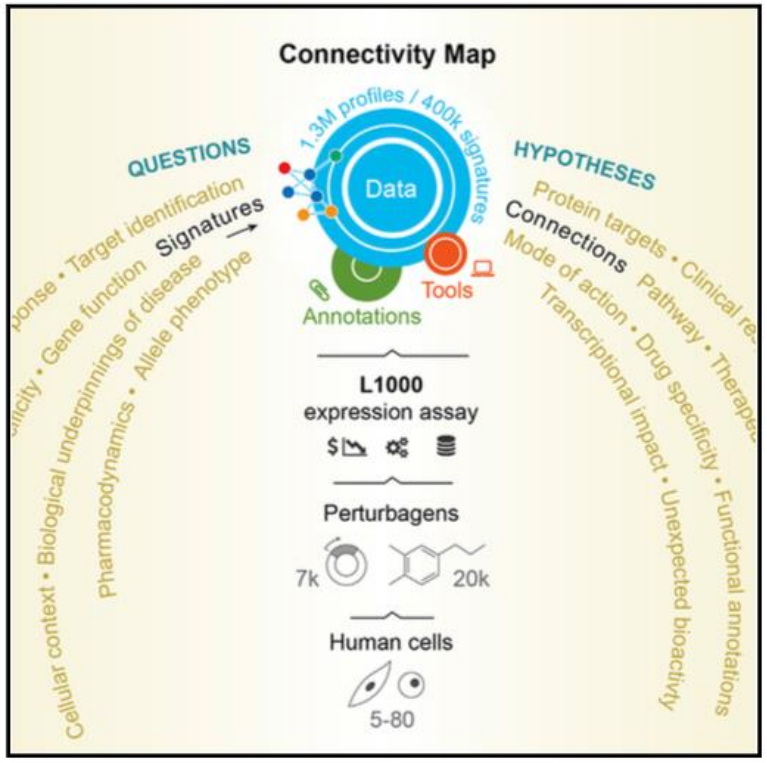
# Next-Generation Connectivity Map

Resource

Cell

## A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles

### Graphical Abstract




### Authors

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, ..., David E. Root, Bang Wong, Todd R. Golub

### Correspondence

golub@broadinstitute.org

### In Brief

The next generation Connectivity Map, a large-scale compendium of functional perturbations in cultured human cells coupled to a gene-expression readout, facilitates the discovery of connections between genes, drugs, and diseases.

Subramanian et al (2017), ***Cell*** 171, 1437 – 1452

# How to make it high-throughput and cheaper?

- Measure a small number of gene expression but still can infer the other genes (genome-wide)
- Need a new technique
- Need a new deconvolution method (computational)

# L1000

## Reduced Representation of Transcriptome

We hypothesized that it might be possible to capture at low cost any cellular state by measuring a reduced representation of the transcriptome. To explore this, we analyzed 12,031 Affymetrix HGU133A expression profiles in the Gene Expression Omnibus (GEO). We used these to identify the optimal number of informative transcripts (*k*), which we term "landmark" transcripts. If *k* was too small, too much information might have been lost, whereas if *k* was too large, sufficient cost reduction compared to the entire transcriptome might be not have been achieved. This analysis showed that 1,000 landmarks were sufficient to recover 82% of the information in the full transcriptome (STAR Methods). The selection of the 1,000 landmarks was done using a data-driven approach rather than selecting transcripts based on prior biological knowledge, as detailed in STAR Methods.

# Defining L1000

## METHOD DETAILS

### Dataset for Landmark selection ($DS_{GEO}$)

We assembled a large, diverse collection of 12,063 gene expression samples profiled on Affymetrix HG-U133A microarrays from the Gene Expression Omnibus (GEO) (Edgar et al., 2002). These data were used to identify the subset of universally informative transcripts to be measured, which we term 'Landmark Genes' (Dataset $DS_{GEO}$).

### Selecting landmark transcripts

As $DS_{GEO}$ contains a non-uniform representation of various aspects of biology (for example certain tumor types such as breast and lung cancer were disproportionately represented), we applied Principal Component Analysis (PCA) as a dimensionality reduction procedure to minimize bias toward any particular lineage or cellular state. In this reduced eigenspace of 386 components (which explained 90% of the variance), cluster analysis was performed to identify tight clusters of commonly co-regulated transcripts. We applied an iterative peel-off procedure to select the centroids (Tseng and Wong, 2005). Specifically, at each iterative step in the tight clustering process, the k-means algorithm with *k* ranging 20-100 was applied repeatedly on 100 independent random subsamples each comprising 75% of the original data. This procedure yielded a consensus matrix that contained the proportion of trials in which a pair of genes were in the same cluster. Thresholding the consensus matrix yielded sets of genes that co-clustered in more than 80% of the trials. The genes belonging to the stable clusters were noted, excluded from the data and the procedure was repeated to identify additional clusters. Because high-dimensional data is challenging to partition into definitive clusters, the advantage of this approach is that gene-gene clusters are derived through the tendency of genes to be grouped together under repeated resampling and hence are more robust to the initialization and cluster size thresholds. Transcripts nominated as landmarks through this process were then tested empirically to assess ability to measure levels accurately in the L1000 assay as described in "Probe and primer design for the L1000 assay" and experimental validation as described in the L1000 reproducibility sections below.

### Evaluating performance of reduced representations of the transcriptome

To simulate performance of measuring a subset of the transcriptome, we asked what number of landmarks (*k*) would optimally recover the observed connections seen in the pilot Connectivity Map dataset based on Affymetrix arrays (Dataset $DS_{CMAP\text{-}AFFX}$). Specifically, prior work indicated that 25 query signatures yielded robust and expected connections to small molecules in the CMap pilot dataset (Table S1). We therefore used those 25 signatures to query the inferred $DS_{CMAP\text{-}AFFX}$ dataset for various values of *k*, counting how often we recovered the connections observed in the original dataset at a comparable rank based on the Kolmogorov-Smirnov statistic. At values of *k* ranging from 100-10,000, we generated an imputed version of $DS_{CMAP\text{-}AFFX}$ using OLS regression (trained on samples from $DS_{GEO}$) with the *k* landmarks as the independent variables, queried it with the benchmark signatures, and assessed the percentage of connections that were recovered.

# Defining L1000

**Baseline expression of landmark genes across a diversity of tissue types**

Our procedure for selecting Landmark Genes was data-driven and the simulations presented above indicate that both the landmark and inferred genes capture relevant information about cell state. However, given a new state, any inference algorithm will only work if a fair number of the landmark genes are expressed in that state. We examined expression across lineage using the Genotype Tissue Expression (GTEx) RNA-seq dataset ($DS_{GTEx\text{-}RNA\text{-}seq}$) of 3,176 patient-derived expression profiles from 30 different tissue types (Figure S1B). We quantified the expression levels of the landmark genes reported in the dataset and observed that at a RPKM threshold of 1 at least 86% of Landmark Genes are expressed in each of the 3,176 samples (with an average of 92% expressed in each sample), and that range of expression is similar across tissue types.

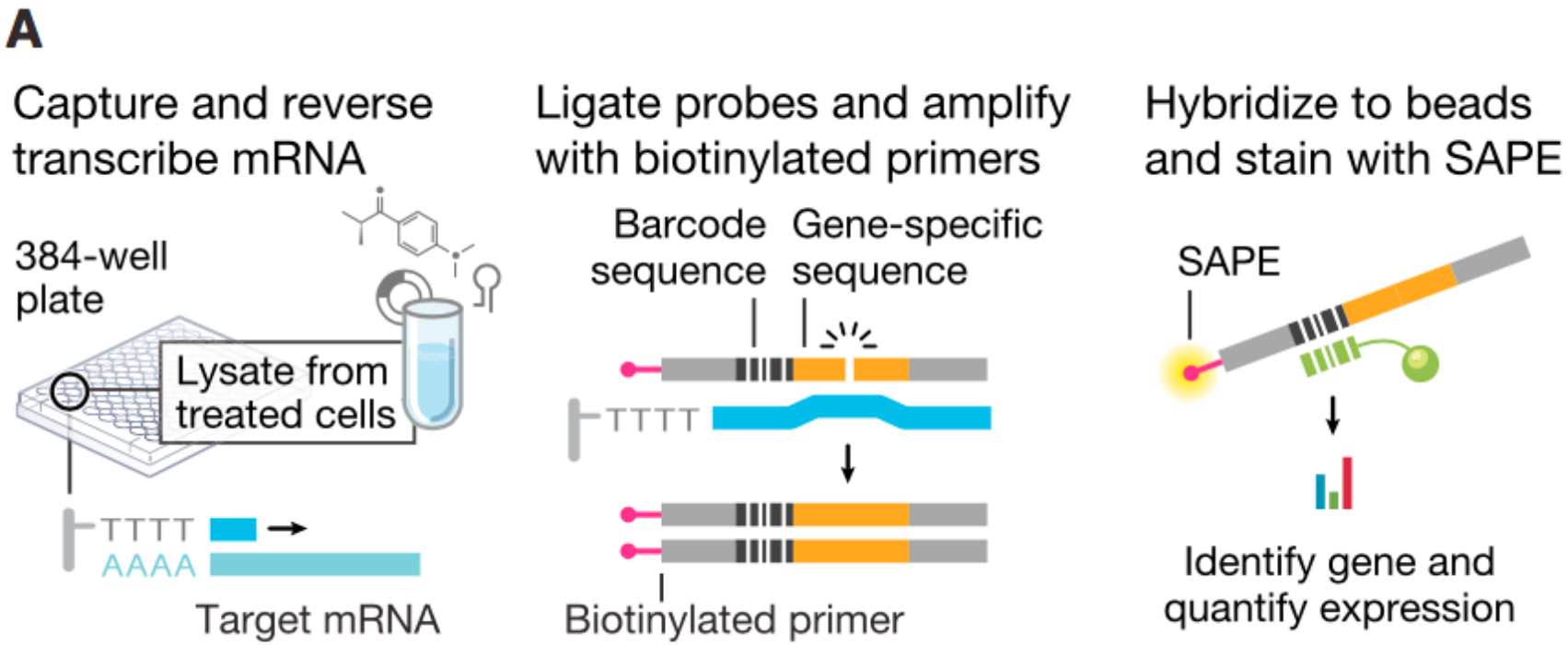
**Functional enrichment analysis of landmark gene content**

Our data-driven procedure suggested genes to include as landmarks based on analysis of the 12,063 sample compendium $DS_{GEO}$. We then asked if genes suggested by this data driven approach were enriched in particular known biological pathways or categories.

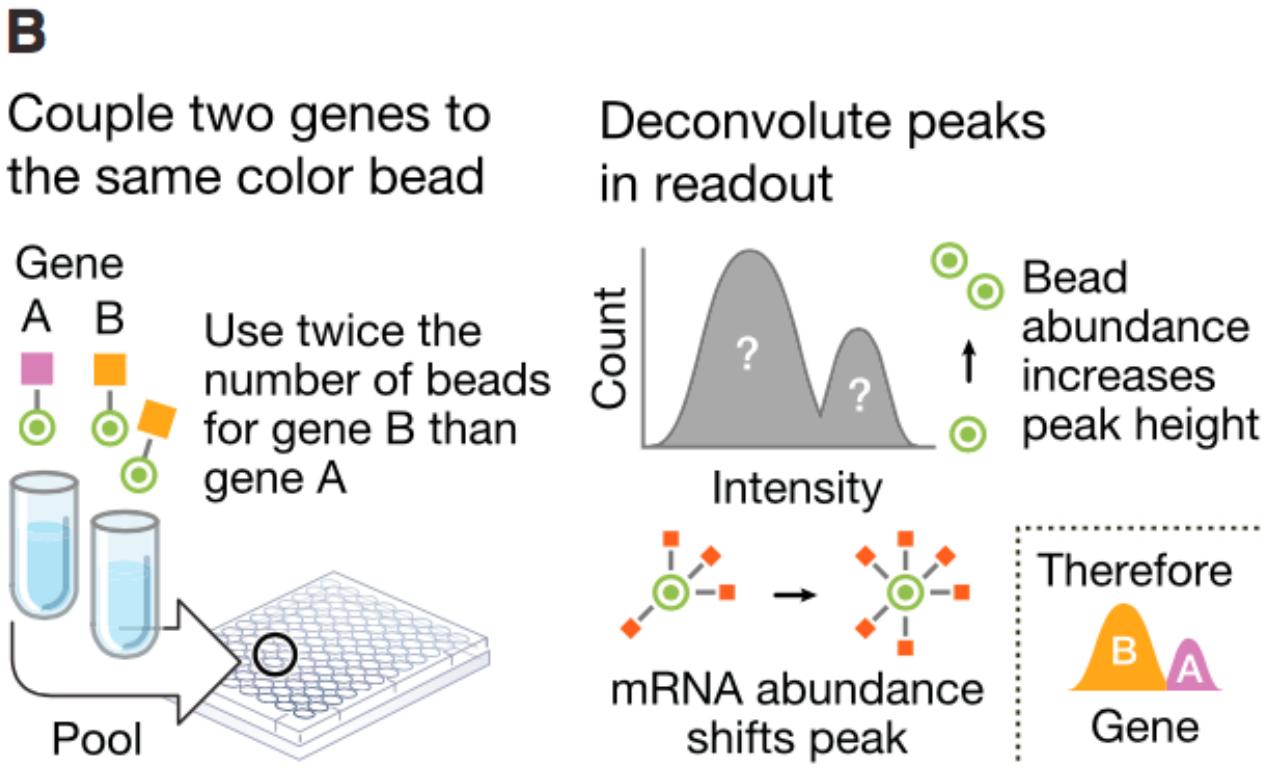
For every landmark gene we accessed from NCBI entrez its current gene description and family assignment. We also annotated every landmark gene with the pathway (as defined in MSigDB) in which it is thought to function (when available). Finally, we looked up its biological/molecular category from Gene Ontology (GO). These annotations were analyzed for functional enrichment to ask if the landmarks, when considered as a set, are dominated by a few functions or if on the whole they map to many different functions. For example, at one extreme the transcriptionally active genes could belong to basic regulatory processes (e.g transcription factors).

To do this analysis we intersected the 978 landmarks with a database of 1,533 gene sets compiled in Gene Ontology using the hypergeometric statistic (gene to GO gene ontology, conditional test for over-representation). We used the R Bioconductor package GOstats (v2.36.0) and the ontology from GO.db (v3.2.2). The results show that while some categories are enriched (e.g ATP binding, nucleoside/nucleotide activity, transcription factor binding, kinase regulator activity) the percentage of the 978 genes that are in any such set is small. While we did observe a number of classes to be enriched in the landmark genes, these categories tend to be generic (e.g., enzyme binding, protein kinase binding, catalytic activity, ATP binding) and/or contain only a small fraction of the landmark genes (e.g., protein kinase binding, which contains 84 of 978 landmarks). Taken together, we did not find any particular functional category dominating the list of landmarks chosen.

# Fig 1A: Overview of ligation-mediated amplification



# Fig 1B: Deconvoluting 1,000 landmark genes using 500 bead colors

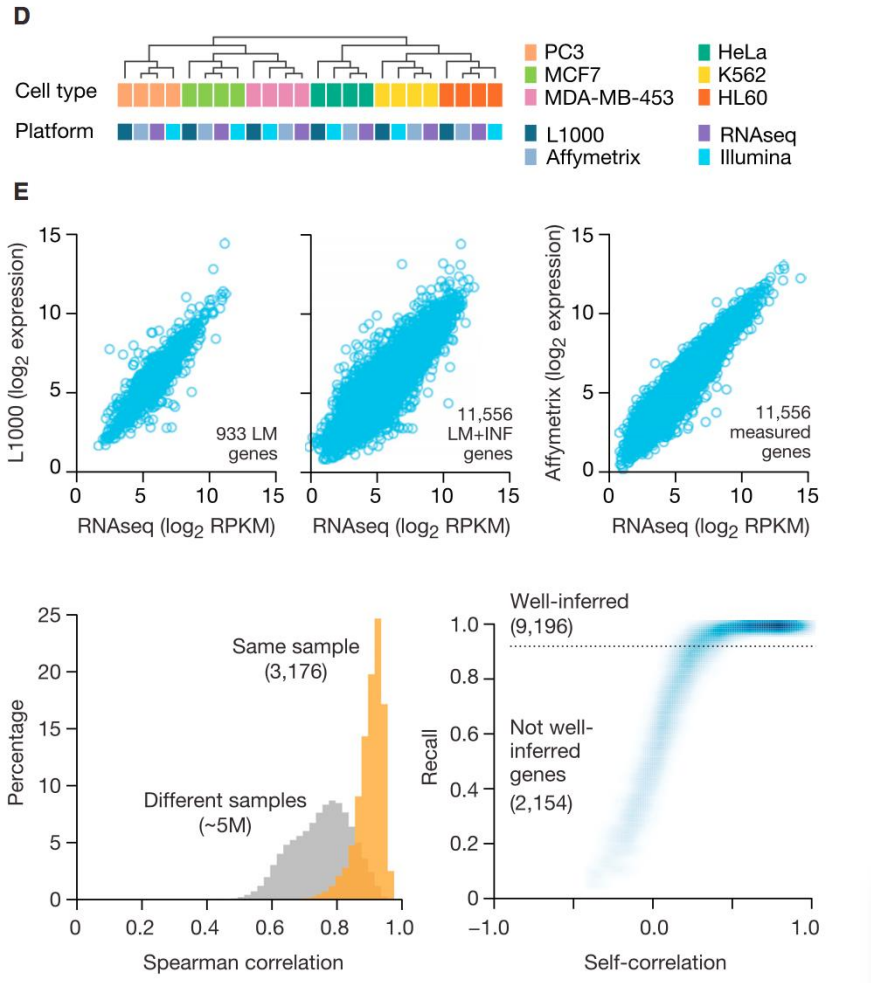


# L1000 to RNA-seq

## Comparison of L1000 to RNA-Seq

RNA-seq has become the standard for gene expression profiling, and thus, we sought to benchmark L1000 against it. We note that while RNA-seq is attractive given its unbiased nature, it suffers from inability to detect non-abundant transcripts without deep sequencing, which results in higher costs. The L1000 platform is hybridization based, thus making the detection of non-abundant transcripts feasible. As an initial assessment of cross-platform performance, mRNA samples from six cell lines were profiled on L1000, by Affymetrix U133A and Illumina BeadChip arrays, and by RNA-seq. Hierarchical clustering of these data grouped samples by cell type, not measurement platform (Figures 1D and 1E [upper panel]).

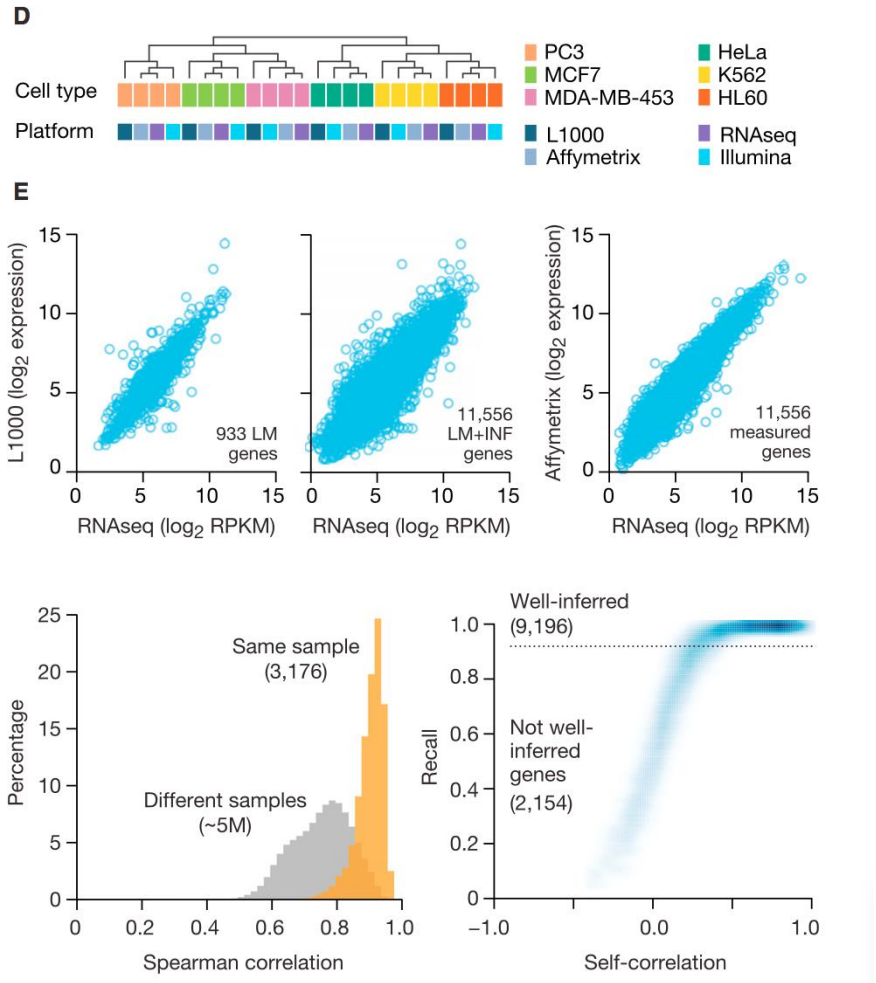
To more extensively compare L1000 to RNA-seq, we analyzed 3,176 samples (previously sequenced by the GTEx Consortium [2015]) profiled on both platforms. This analysis showed that cross-platform similarity was high (median self-correlation, 0.84), with a right-shifted distribution compared to non-self correlations (Figure 1E [lower left panel]). Recall analysis similarly showed that 98% of samples had a sample recall >99% (indicating 99th percentile) (STAR Methods). Taken together, these results indicate a strong degree of similarity of profiles across L1000 and RNA-seq platforms.



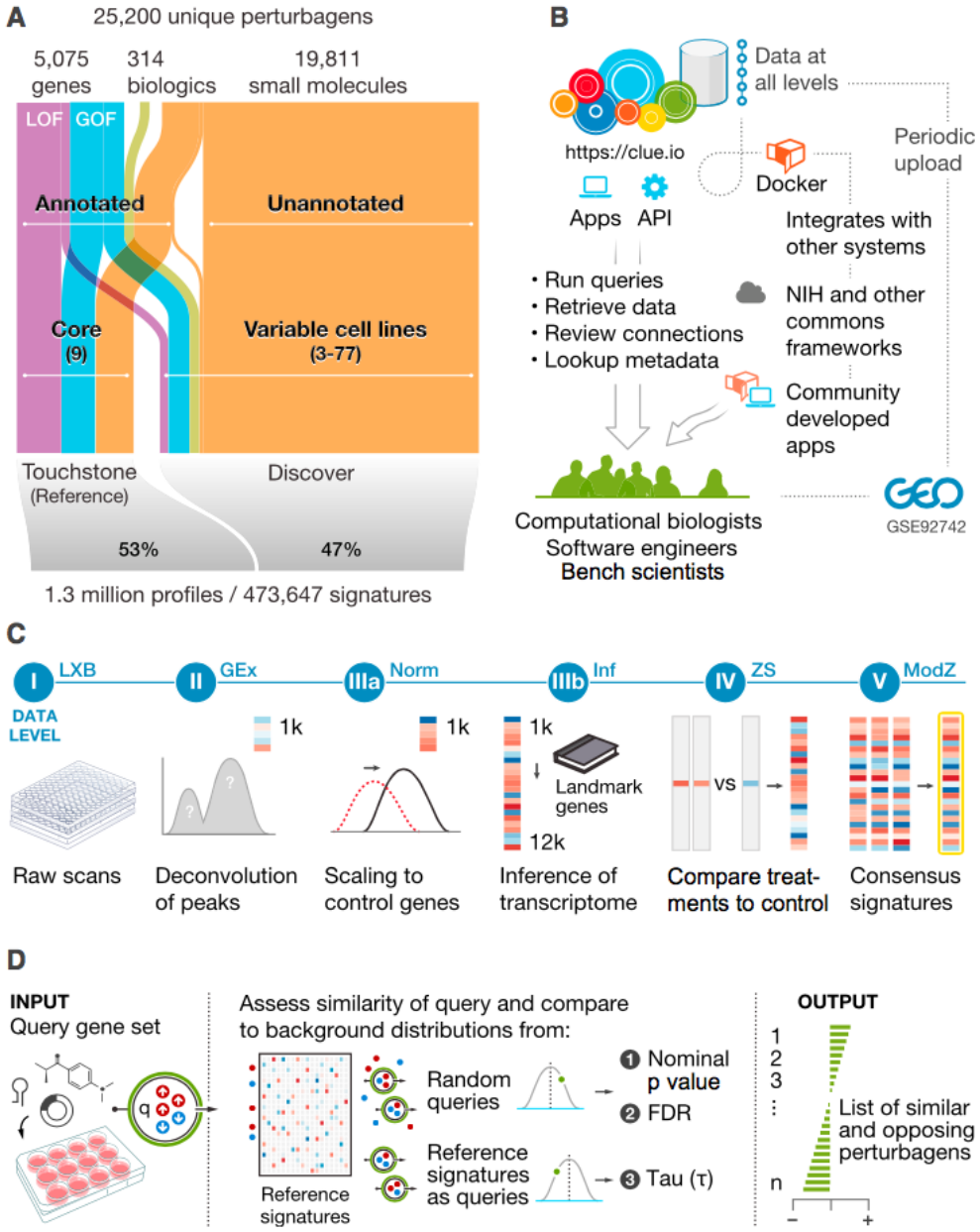
# L1000 to RNA-seq

**Inferring Gene Expression from L1000 Landmarks**

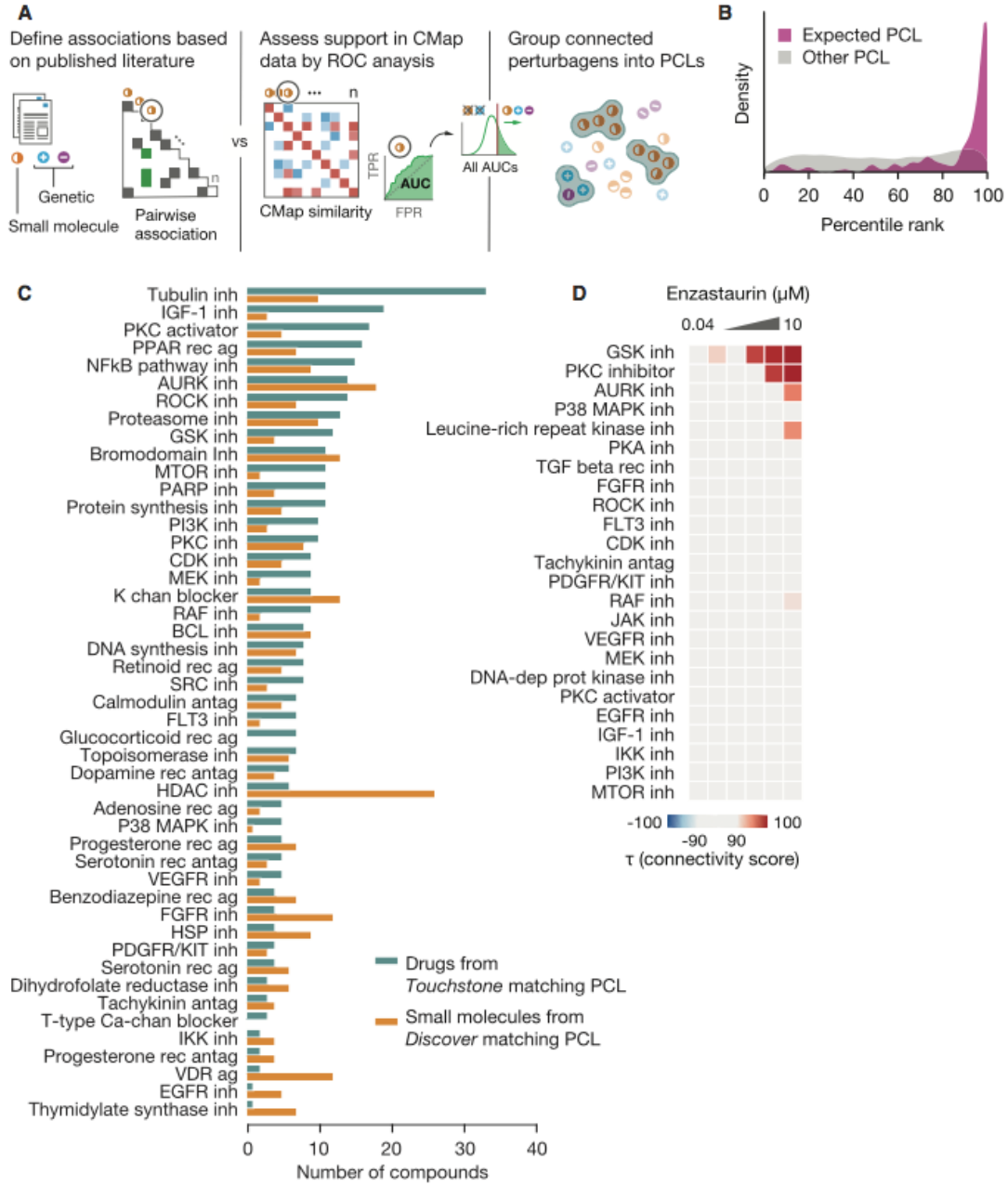
Using 8,555 RNA-seq samples (dataset $DS_{GTEx\text{-}rnaseq}$) as an independent test set, we used landmark transcript measurements to infer the remainder of the transcriptome. As a test of inference accuracy, we analyzed gene-level recall ($R_{gene}$) for each of the inferred genes and assessed performance by comparing the result to a null distribution of correlations between all inferred transcripts and all measured transcripts. This analysis showed that inference was accurate (defined as $R_{gene} \geq 0.95$) for 9,196 of the 11,350 inferred genes (81%). When combined with the 978 measured landmarks, the L1000 platform thus measures or infers with high fidelity 83% of transcripts, but yields poor inference for 17% (Figure 1E [lower panel right] and Table S3). Inferences for these 17% were therefore not used in any of the analyses that follow.



# Fig 2: L1000 Dataset Coverage, Signature Generation & Data Access



# Fig 4. Reference Perturbagen Classes for CMap Discovery



# Query: Computing Similarities

### Query methodology

The fundamental unit of CMap analysis is the query. A query ($q$) consists of a set of genes corresponding to any biological state of interest. Each gene in the query carries a sign indicating whether it is upregulated or downregulated. Thus each query yields a pair of mutually exclusive gene lists ($q_{up}$, $q_{down}$). The query is compared to each signature in the CMap reference database (*Touchstone*) using the similarity metric described below to assess connectivity or the degree to which the upregulated query genes ($q_{up}$) appear toward the top of the rank-ordered signature and the downregulated query genes ($q_{down}$) appear toward the bottom of the signature (positive connectivity) or vice-versa (negative connectivity). The result of a query is a rank ordered list of CMap signatures ordered by their connectivity scores.

### Computing similarities - Weighted Connectivity Score (*WTCS*)

The weighted connectivity score (*WTCS*) represents a non-parametric similarity measure based on the weighted Kolmogorov-Smirnov enrichment statistic (*ES*) described previously (Subramanian et al., 2005). *WTCS* is a composite, bi-directional version of *ES*. For a given query gene set pair ($q_{up}$, $q_{down}$) and a reference signature $r$, *WTCS* is computed as follows:

$$w_{q,r} = \begin{cases} (ES_{up} - ES_{down})/2, & \text{if } \operatorname{sgn}(ES_{up}) \neq \operatorname{sgn}(ES_{down}) \\ 0, & \text{otherwise} \end{cases}$$

Where $ES_{up}$ is the enrichment of $q_{up}$ in $r$ and $ES_{down}$ is the enrichment of $q_{down}$ in $r$. *WTCS* ranges between −1 and 1. It will be positive for signatures that are positively related and negative for those that are inversely related, and near zero for signatures that are unrelated. A null (0) score is assigned for cases when both $ES_{up}$ and $ES_{down}$ are the same sign.

### Normalization of Connectivity Scores

To allow for comparison of connectivity scores across cell types and perturbation types, the scores are normalized to account for global differences in connectivity that might occur across these covariates. Given a vector of *WTCS* values $w$ resulting from a query, we normalize the values within each cell line and perturbagen type to obtain normalized connectivity scores (*NCS*) as follows:

$$NCS_{c,t} = \begin{cases} w_{c,t} \Big/ \mu^{+}_{c,t} & \text{if } \operatorname{sgn}(w_{c,t}) > 0 \\ w_{c,t} \Big/ \mu^{-}_{c,t} & \text{otherwise} \end{cases}$$

where $NCS_{c,t}$, $w_{c,t}$, $\mu^{+}_{c,t}$ and $\mu^{-}_{c,t}$ are the normalized connectivity scores, raw weighted connectivity scores, and signed means of the raw weighted connectivity scores (the mean of positive and negative values evaluated separately) within the subset of *Touchstone* signatures corresponding to cell line $c$ and perturbagen type $t$, respectively.

Overall, this procedure is similar to that used in Gene Set Enrichment Analysis, with the addition of bidirectional gene sets (i.e., up and down) as queries.

# Connectivity Map Score ($\tau$)

## Connectivity Map Score

Tau ($\tau$) compares an observed enrichment score to all others in a reference database. In principle, $\tau$ can be computed by comparison to scores from any database of reference signatures, and the most common approach is to generate a null distribution by random permutation. However, a more stringent test that avoids having to make assumptions regarding the complex correlation structure of gene expression data is to use a compendium of diverse, biologically relevant perturbational signatures, such as those in CMap-L1000v1, as it is these reference signatures against which any novel connection must compete. Thus, query results are scored with $\tau$ as a standardized measure ranging from $-100$ to 100; a $\tau$ of 90 indicates that only 10% of reference perturbations showed stronger connectivity to the query. Because the reference is fixed, $\tau$ can be used to compare results across queries - a connection with a significant p value and FDR but low $\tau$ would suggest a highly promiscuous relationship whose connections are not unique.

## Calculating $\tau$

While meaningful comparisons can be made between the *NCS* values of reference signatures with respect to query *q*, it is also useful to assess if the connectivity between *q* and a particular signature *r* is significantly different from that observed between *r* and other queries. This is done by comparing each observed *NCS* value $ncs_{q,r}$ between the query *q* and a reference signature *r* to a distribution of *NCS* values representing the similarities between a reference compendium of queries ($Q_{ref}$) and *r*. This procedure results in a standardized measure we refer to as Tau ($\tau$) that ranges from $-100$ to $+100$ and represents the percentage of queries in $Q_{ref}$ with a lower $|NCS|$ than $|ncs_{q,r}|$, adjusted to retain the sign of $ncs_{q,r}$ :

$$\tau_{q,r} = \text{sgn}(ncs_{q,r}) \frac{100}{N} \sum_{i=1}^{N} \left[ |ncs_{i,r}| < |ncs_{q,r}| \right]$$

where $ncs_{q,r}$ is the normalized connectivity score for signature *r* w.r.t query *q*, $ncs_{i,r}$ is the normalized connectivity score for signature *r* relative to the i-th query in $Q_{ref}$ and *N* is the number of queries in $Q_{ref}$. Our standard practice is that $Q_{ref}$ be comprised of queries obtained from exemplar signatures of *Touchstone* perturbagens that match the cell line and perturbation type of signature *r*. In principle any arbitrary compendium of gene sets (as long as they are large enough) could be used.

## Summarization Across Cell Lines

When examining query results, it is often convenient to obtain a perturbagen-centric measure of connectivity that summarizes the results observed in individual cell types. This can be particularly helpful when searching for connections that persist across cell lines or when one is unsure which cell line to examine. Given a vector of normalized connectivity scores for perturbagen *p*, relative to query *q*, across all cell lines in which *p* was profiled, a cell-summarized connectivity score is obtained using a maximum quantile statistic:

$$NCS_{c,t} = \begin{cases} Q_{hi}(\boldsymbol{ncs}_{p,c}) & \text{if } |Q_{hi}(\boldsymbol{ncs}_{p,c})| \geq |Q_{lo}(\boldsymbol{ncs}_{p,c})| \\ Q_{lo}(\boldsymbol{ncs}_{p,c}) & \text{otherwise} \end{cases}$$

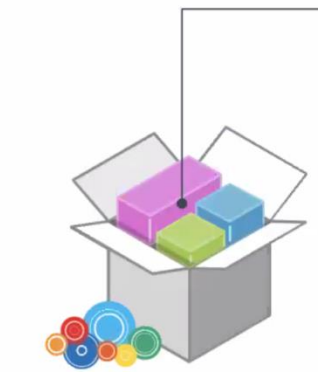
where $\boldsymbol{ncs}_{p,c}$ is a vector of normalized connectivity scores for perturbagen *p*, relative to query *q*, across all cell lines in which *p* was profiled, and $Q_{hi}$ and $Q_{lo}$ are upper and lower quantiles respectively. This procedure compares the $Q_{hi}$ and $Q_{lo}$ quantiles of $\boldsymbol{ncs}_{p,c}$ and retains whichever is of higher absolute magnitude. Thus, maximum quantile is more sensitive to signal in a subset of the cell lines than measures of central tendency such as mean or median. In the analyses presented here, we used

$$Q_{hi} = 67, \; Q_{lo} = 33$$

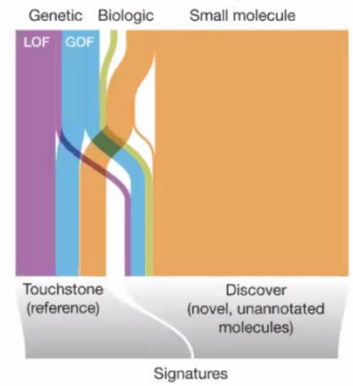
# 3,000,000 profiles

## The final LINCS dataset

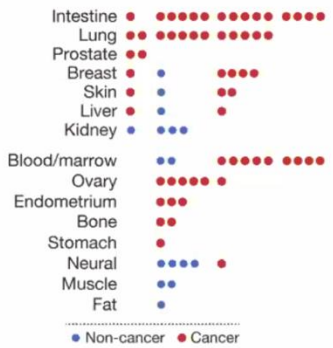
L1000 — >3M profiles



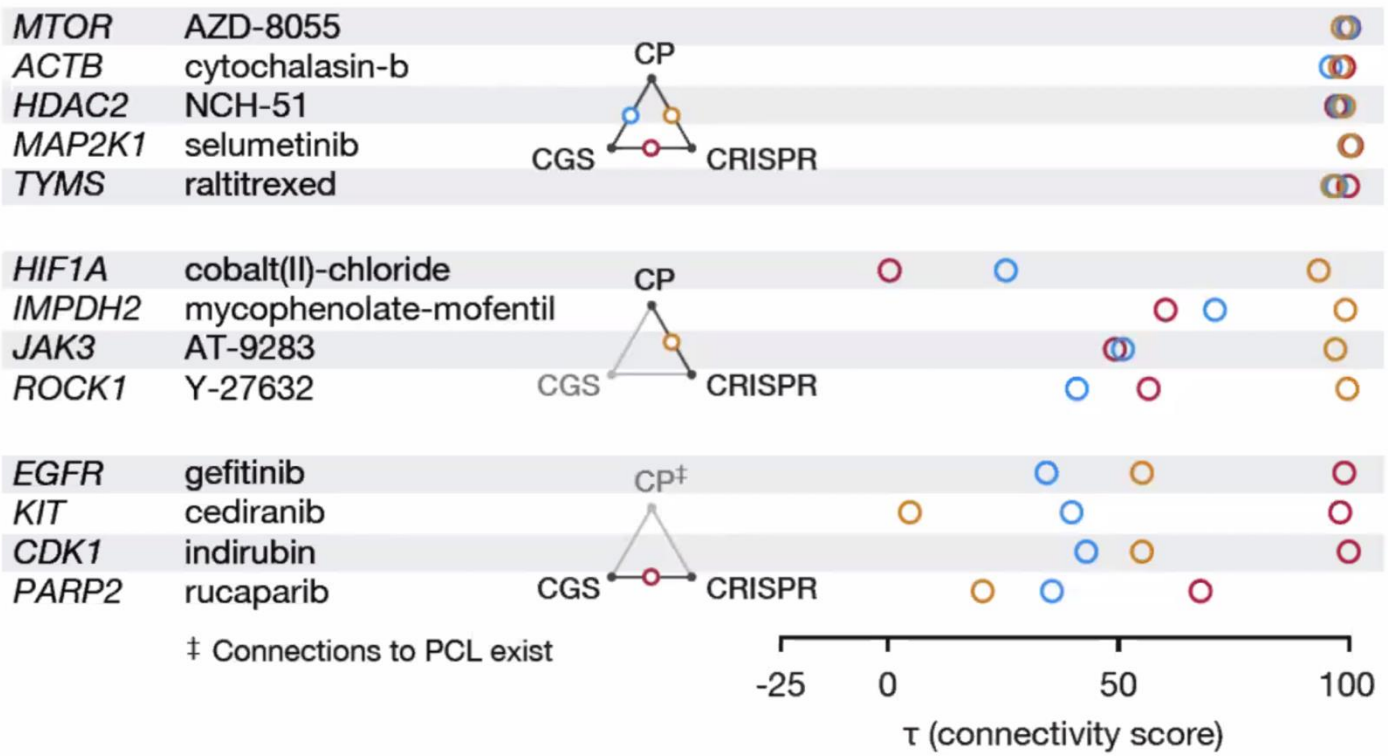

92,648 perturbations
Incl. 25,590 compounds




253 contexts
Incl. 22 primary cells



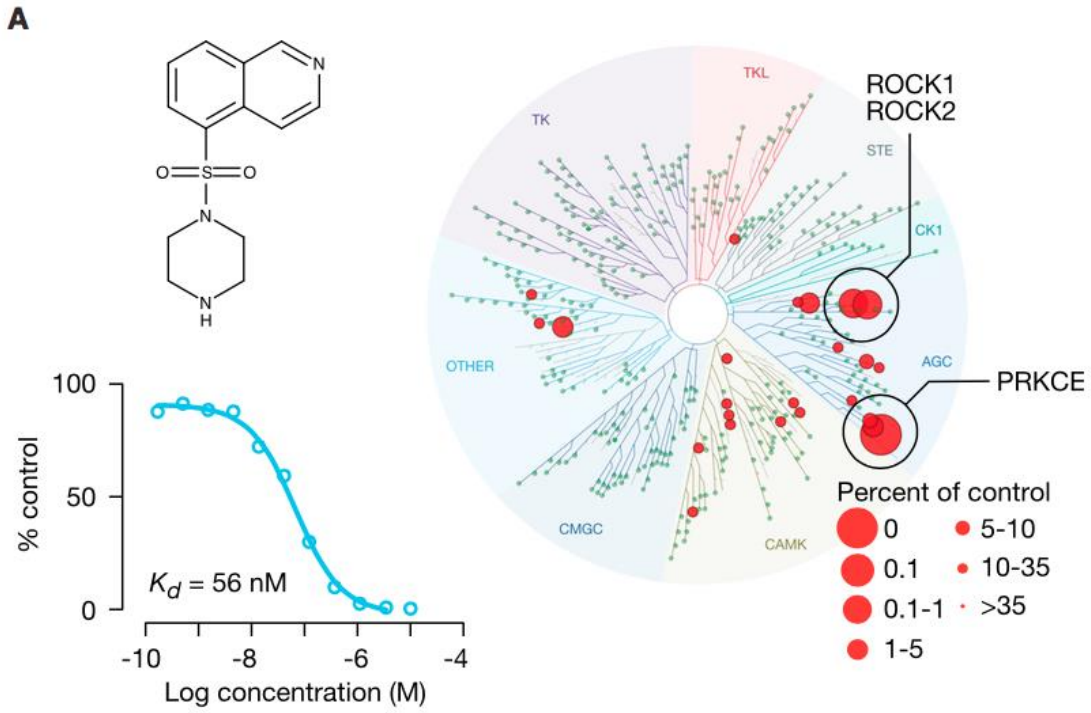
# Genetic vs. chemical perturbation to inform MoA



# Applications

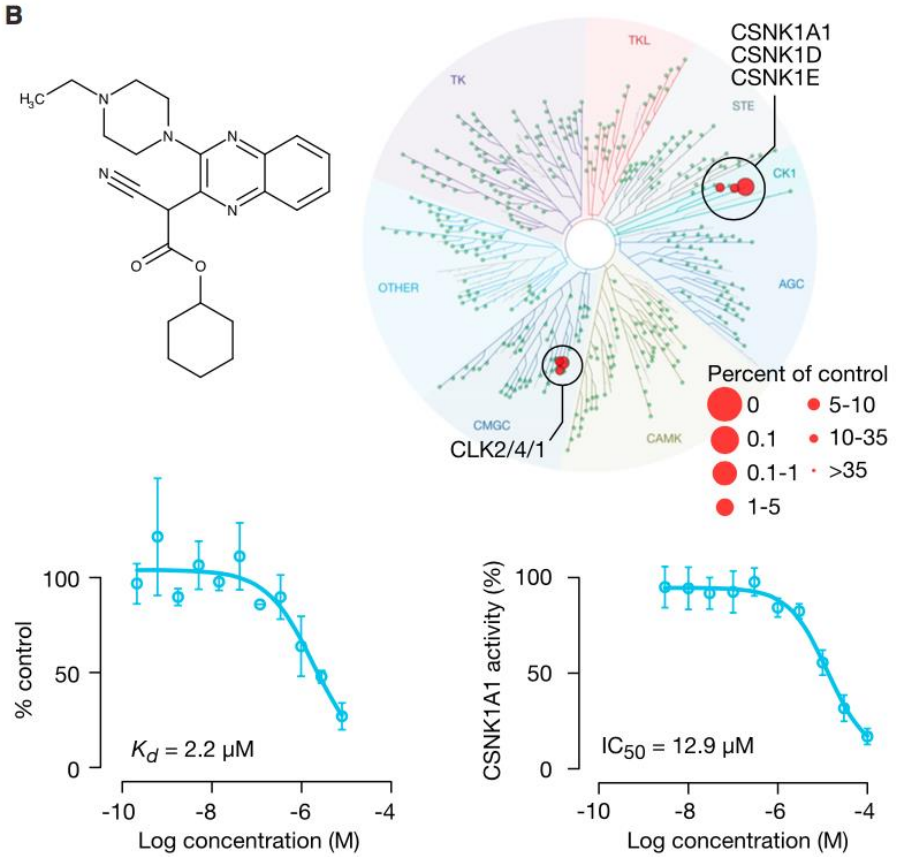
# Fig 6A: Discovery of MOA

Query: Find compounds that induce the similarity gene expression profiles as query signature.
Unannotated compound BRD-2751 showed strong connectivity to the Rho-associated protein kinase (ROCK) PCL, suggesting that it might in fact be a ROCK inhibitor.

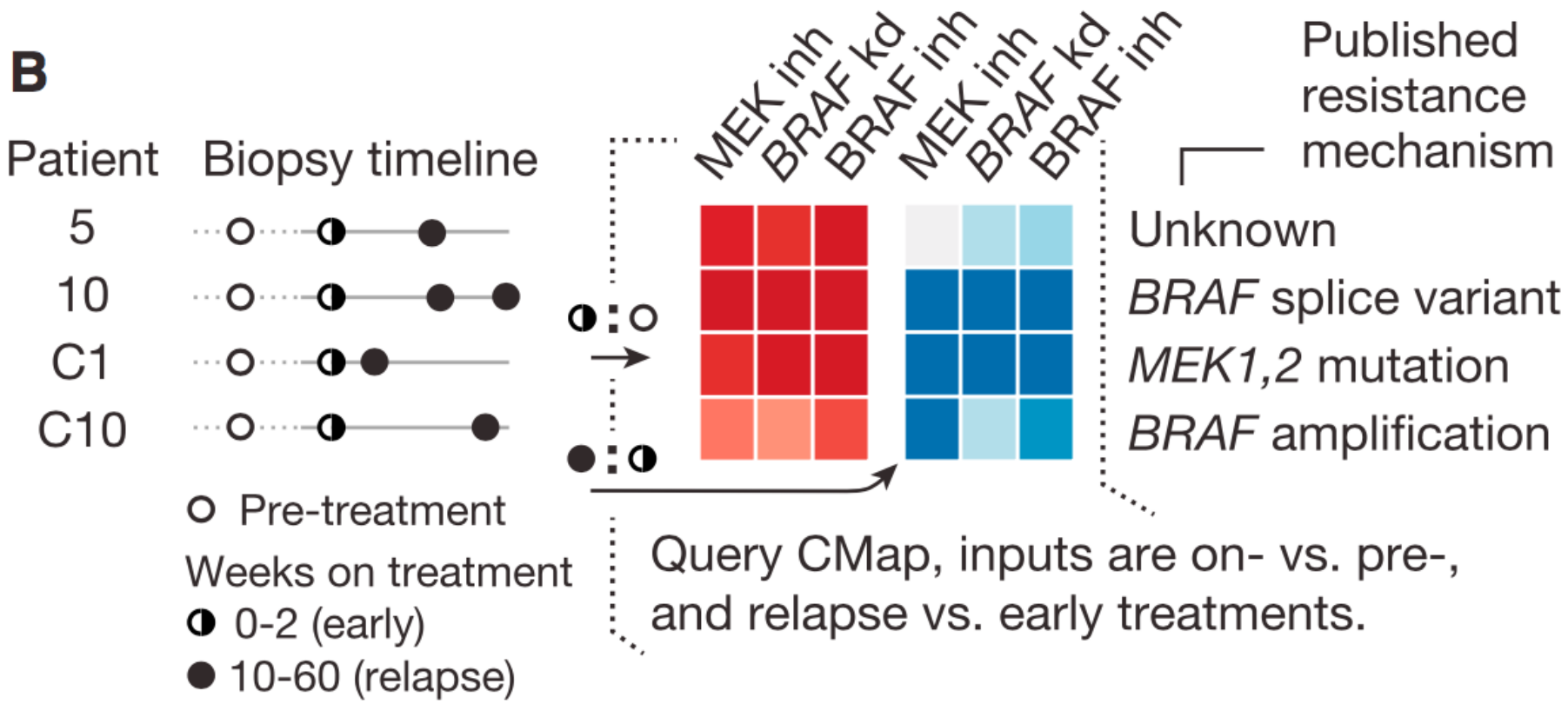


# Fig 6B: Discovery of Selective Compound

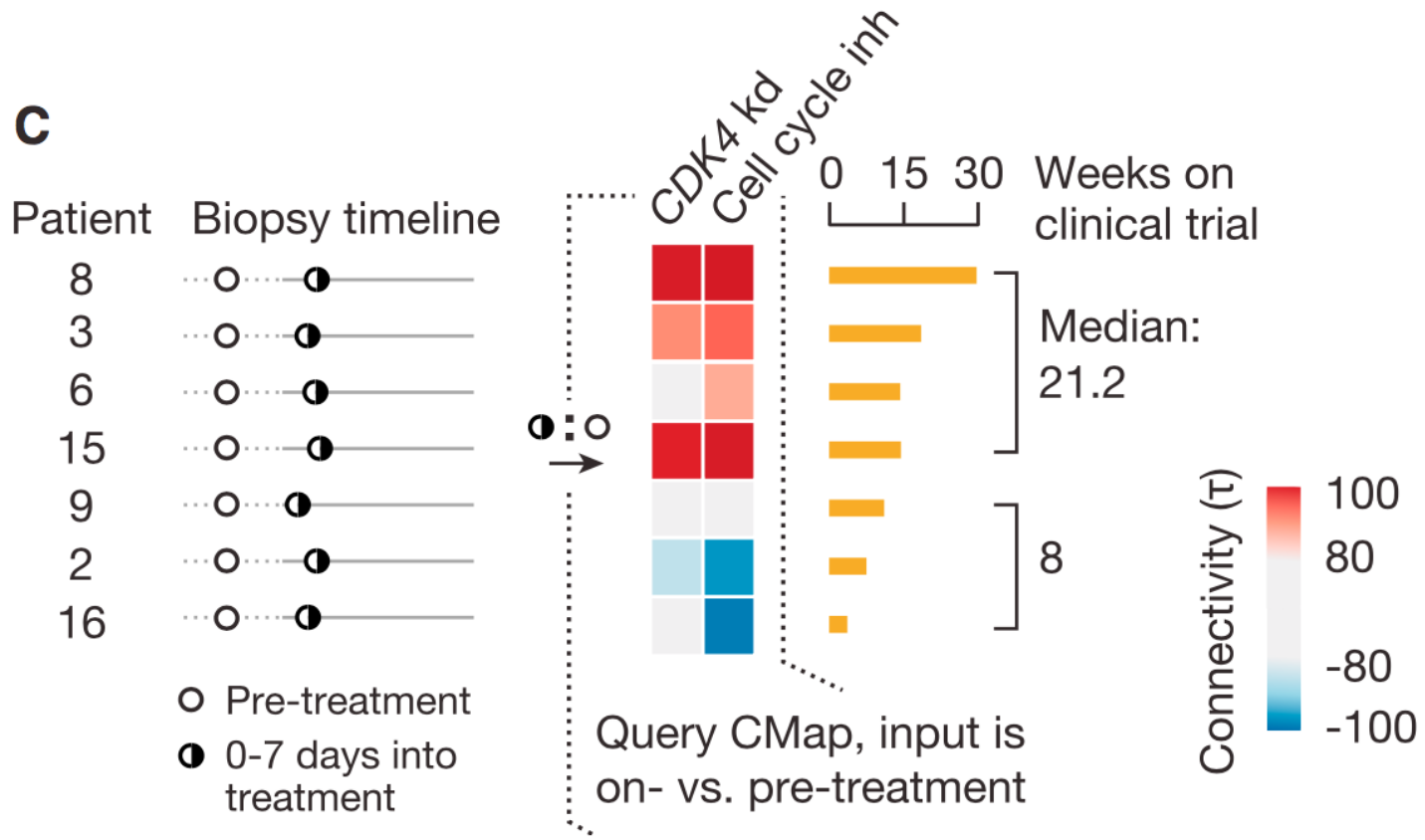
Query: Find compounds that induce the similarity gene expression profiles as Loss of function (shRNAs CSKN1A1).
Results: One unannotated compound BRD-1868 showed strong connectivity to CSNK1A1 knock-down in two cell types.



# Fig 7B: Connecting Patients data to explain resistance mechanisms



# Fig 7C: Connecting Patients data to predict therapeutic efficacy



[clue.io]

clue.io

Update

CLUE

Tools Projects Developer Partnering Help | Log in

ConnectivityMap

BROAD INSTITUTE

Unravel biology with the world's largest perturbation-driven gene expression dataset.

Start exploring the data by using the text-box on this page to look up perturbagens of interest in Touchstone. To see the suite of tools, including apps to query your gene expression signatures and analyze resulting connections, click on Tools in the menu bar.

> TYPE COMPOUND, GENE, MoA, OR PERTURBAGEN CLASS TO SEE OVERVIEW
> TYPE A SLASH CHARACTER "/" TO SEE LIST OF COMMANDS

DATA VERSION: Beta / SOFTWARE VERSION: 1.1.1.43

**PRESENTATIONS AND VIDEOS FROM THE DECEMBER 2020 WORKSHOP CAN BE FOUND BY CLICKING ON THE LINK BELOW**

https://clue.io/workshop-2020/workshop-resources.

https://clue.io

## Data and Tools

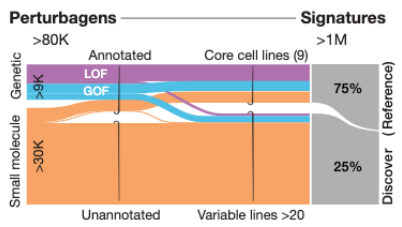
We are excited to announce the release of the updated CMap LINCS gene expression resource. This release is an expansion upon the previous 2017 data release and contains >3M gene expression profiles and >1M replicate-collapsed signatures.

1. These data are available for download from the LINCS data releases app as well as the from the clue data library .
2. The data can be queried with external gene sets using the clue query app
3. We also provide a web application for querying the metadata
4. And a python library for accessing the data programmatically

In addition, we provide the following tools to help facilitate data access and use:

1. Code libraries for accessing and analyzing CMap data
2. Notebooks that illustrate common modes of data access and analysis
3. Docker containers for running common analysis algorithms

**Please note that these data and tools are released as a beta version and will likely be subject to change as minor updates are made.**




BROAD INSTITUTE PATTERN SEE & UNDERSTAND DATA

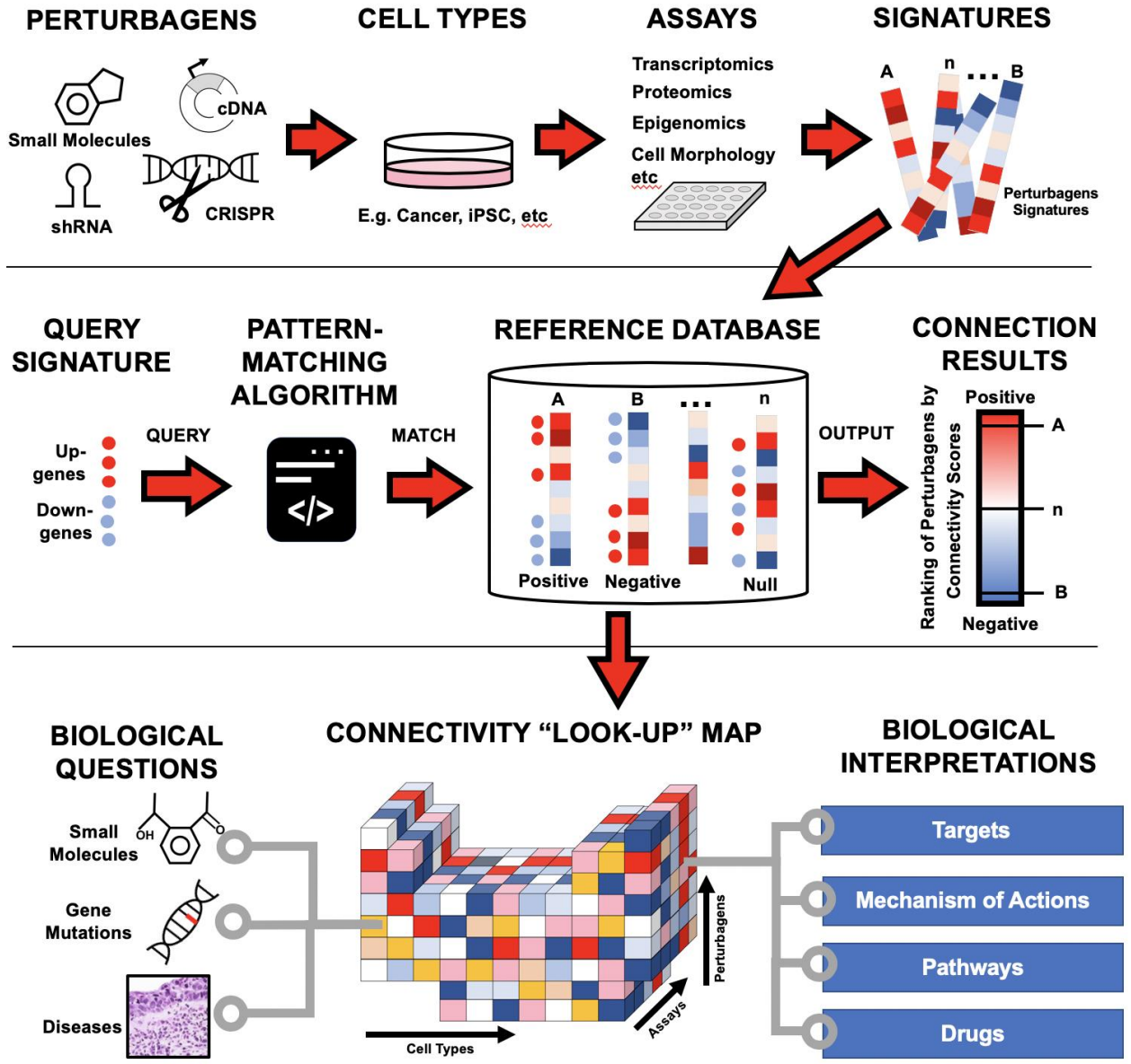
About
CMap
CLUE
Team
Careers
Acknowledgements

Support
Contact Us
Connectopedia: Clue Knowledge Base
Office hours



Branch: master Build Date: Fri Jul 02 2021 Revision: 04cb1529d4eb471515446342 9e0c2c585f6d2d2e Short: 04cb152

# Beyond Gene Expression-based CMap




(Lee, Cheong, Tan, *Adv. Mol. Path.* 2019)

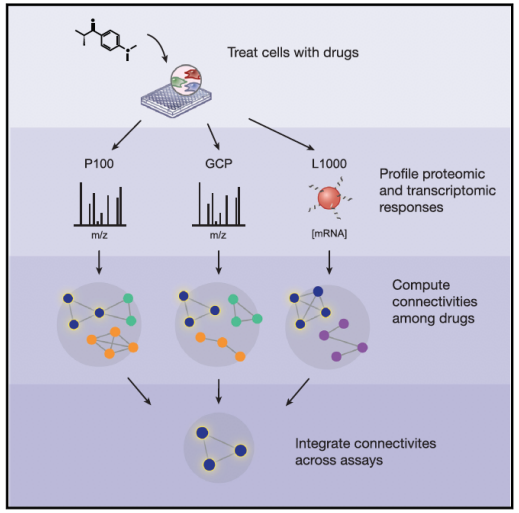
# Proteomics & Chromatin Connectivity Map

Article

Cell Systems

## A Library of Phosphoproteomic and Chromatin Signatures for Characterizing Cellular Responses to Drug Perturbations

### Graphical Abstract




### Authors

Lev Litichevskiy, Ryan Peckner, Jennifer G. Abelin, ..., Michael J. MacCoss, Li-Huei Tsai, Jacob D. Jaffe

### Correspondence

jjaffe@broadinstitute.org

### In Brief

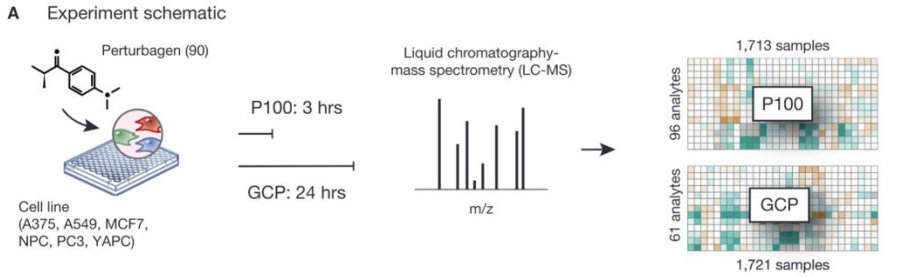
A large compendium of cellular responses to drugs as profiled through proteomic assays of phosphosignaling and histone modifications reveals cellular responses that transcend lineage, discovers unexpected associations between drugs, and recognizes therapeutic hypotheses for treatment of multiple myeloma and acute lymphocytic leukemia.

### Highlights

- First-of-its-kind public resource of proteomic responses to systematic drug treatment
- Profiling of phosphosignaling and chromatin states induced by 90 drugs in 6 cell lines
- Extends Connectivity Map concept to proteomics and integrates with transcriptomics
- Enables recognition of cell type-specific activities and therapeutic opportunities

Litichevskiy et al., 2018, Cell Systems 6, 424–443
April 25, 2018 © 2018 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2018.03.012

CellPress




ACS chemical biology

Articles

Cite This: ACS Chem. Biol. 2020, 15, 140−150

## A Proteomic Connectivity Map for Characterizing the Tumor Adaptive Response to Small Molecule Chemical Perturbagens

Zhenzhen Zi,[†] Yajie Zhang,[†] Peng Zhang,[§,∥] Qing Ding,[†] Michael Chu,[†] Yiwen Chen,[§] John D. Minna,[‡] and Yonghao Yu*[,†]

[†]Department of Biochemistry, UT Southwestern Medical Center, Dallas, Texas 75390, United States

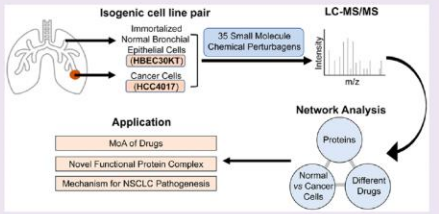
[‡]Hamon Center for Therapeutic Oncology Research, Departments of Internal Medicine and Pharmacology, Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, Texas 75390, United States

[§]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States

[∥]Key Laboratory of RNA Biology, Center for Big Data Research in Health, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

Supporting Information

**ABSTRACT:** A powerful means to understand the cellular function of corrupt oncogenic signaling programs requires perturbing the system and monitoring the downstream consequences. Here, using a unique pair of non-small cell lung cancer (NSCLC)/normal lung epithelial patient-derived cell lines (HCC4017/HBEC30KT), we systematically interrogated the remodeling of the NSCLC proteome upon treatment with 35 chemical perturbagens targeting a diverse array of mechanistic classes. HCC4017 and HBEC30KT cells differ significantly in their proteomic response to the same compound treatment. Using protein covariance analyses, we identified a large number of functional protein networks. For example, we found that a poorly studied protein, C5orf22, is a novel component of the WBP11/PQBP1 splicing complex. Depletion of C5orf22 leads to the aberrant splicing and expression of genes involved in cell growth and immunomodulation. In summary, we show that by systematically measuring the tumor adaptive responses at the proteomic level, an understanding could be generated that provides critical circuit-level biological insights for these pharmacologic perturbagens.



# Cell Morphology Connectivity Map

PROTOCOL

## Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes

Mark-Anthony Bray[1], Shantanu Singh[1], Han Han[2], Chadwick T Davis[2], Blake Borgeson[2], Cathy Hartland[3], Maria Kost-Alimova[3], Sigrun M Gustafsdottir[3], Christopher C Gibson[2] & Anne E Carpenter[1]

[1]Imaging Platform, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [2]Recursion Pharmaceuticals, Salt Lake City, Utah, USA. [3]Center for the Science of Therapeutics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. Correspondence should be addressed to C.C.G. (chris.gibson@recursionpharma.com) or A.E.C. (anne@broadinstitute.org).

**In morphological profiling, quantitative data are extracted from microscopy images of cells to identify biologically relevant similarities and differences among samples based on these profiles. This protocol describes the design and execution of experiments using Cell Painting, which is a morphological profiling assay that multiplexes six fluorescent dyes, imaged in five channels, to reveal eight broadly relevant cellular components or organelles. Cells are plated in multiwell plates, perturbed with the treatments to be tested, stained, fixed, and imaged on a high-throughput microscope. Next, an automated image analysis software identifies individual cells and measures ~1,500 morphological features (various measures of size, shape, texture, intensity, and so on) to produce a rich profile that is suitable for the detection of subtle phenotypes. Profiles of cell populations treated with different experimental perturbations can be compared to suit many goals, such as identifying the phenotypic impact of chemical or genetic perturbations, grouping compounds and/or genes into functional pathways, and identifying signatures of disease. Cell culture and image acquisition takes 2 weeks; feature extraction and data analysis take an additional 1–2 weeks.**

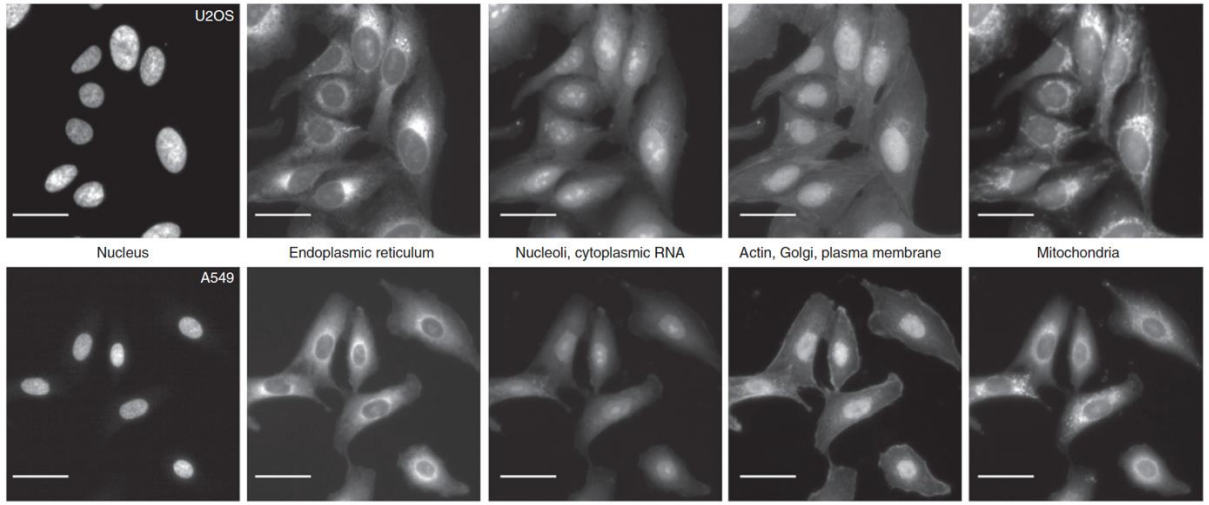



**Figure 1** | The Cell Painting assay in U2OS and A549 cells. The columns display the five channels imaged in the Cell Painting assay protocol (left to right) as imaged using the ImageXpress XLS microscope: Hoechst 33342 (DNA), concanavalin A (endoplasmic reticulum), SYTO 14 (nucleoli and cytoplasmic RNA), phalloidin (actin) and WGA (Golgi and plasma membrane), and MitoTracker Deep Red (mitochondria). Scale bars, 20 µm. See **Table 1** for additional details about the stains and channels imaged.

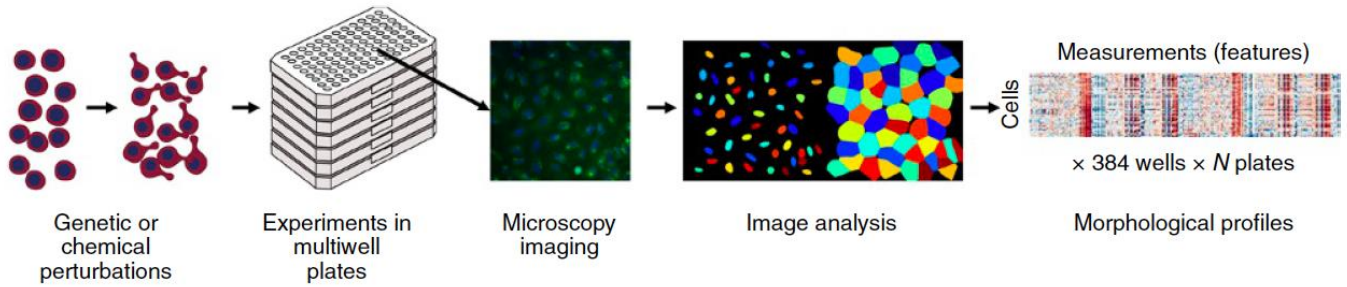



**Figure 2** | Overview of the strategy of morphological profiling using an image-based assay. After perturbing, staining and imaging cells, the open-source software CellProfiler is used to extract ~1,500 morphological features of each cell. The collection of features is known as a *profile*: it reflects the phenotypic state of the cells in that sample, and it can be compared with other profiles to make inferences.

# Cell Morphology Connectivity Map

MOLECULAR ONCOLOGY 1 (2007) 84–96

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/molonc

## The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression

*Paraic A. Kenny*[a,1], *Genee Y. Lee*[a,1], *Connie A. Myers*[a,2], *Richard M. Neve*[a], *Jeremy R. Semeiks*[a], *Paul T. Spellman*[a], *Katrin Lorenz*[a,3], *Eva H. Lee*[a], *Mary Helen Barcellos-Hoff*[a], *Ole W. Petersen*[b], *Joe W. Gray*[a], *Mina J. Bissell*[a,*]

[a]*Life Sciences Division, Lawrence Berkeley National Laboratory, University of California, One Cyclotron Road, MS 977-225A, Berkeley, CA 94720, USA*

[b]*Structural Cell Biology Unit, Institute of Medical Anatomy, The Panum Institute, University of Copenhagen, DK-2200 Copenhagen N, Denmark*

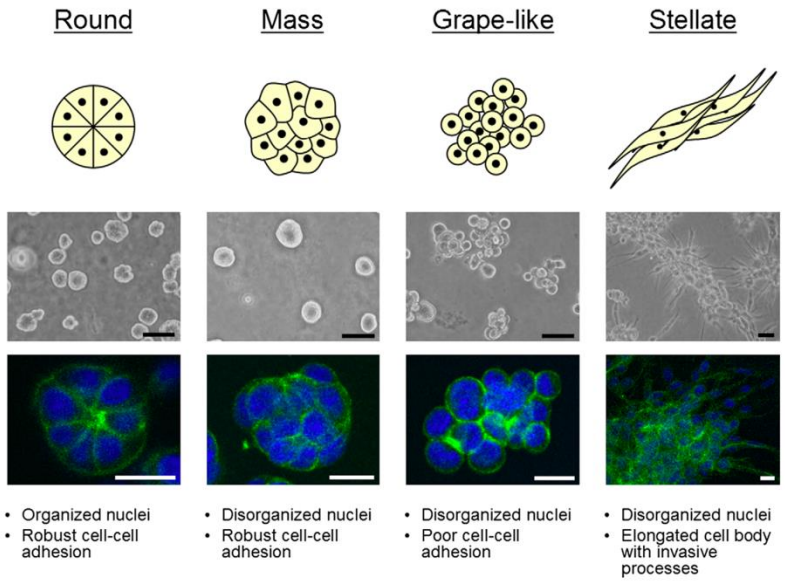



Figure 1 – Breast cell line colony morphologies in 3D culture fall into four distinct groups. A panel of 25 breast cell lines were cultured in three-dimensions and grouped into four distinct morphologies. A schematic and key descriptors of each morphology is shown in addition to phase contrast and F-actin and nuclear fluorescence images of representative cell lines of each morphology: for Round, S1 is shown; Mass, BT-474; Grape-like, SK-BR-3; and Stellate, MDA-MB-231. Scale bars: phase contrast, 50 μm; fluorescence, 20 μm.

*Article*

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

## A chemical–genetic interaction map of small molecules using high-throughput imaging in cancer cells

Marco Breinig[1,2,†], Felix A Klein[3,†], Wolfgang Huber[3,*] & Michael Boutros[1,2,**]

Mol Syst Biol. 2015 Dec 23;11(12):846.

**Abstract**

**Small molecules often affect multiple targets, elicit off-target effects, and induce genotype-specific responses. Chemical genetics, the mapping of the genotype dependence of a small molecule's effects across a broad spectrum of phenotypes can identify novel mechanisms of action. It can also reveal unanticipated effects and could thereby reduce high attrition rates of small molecule development pipelines. Here, we used high-content screening and image analysis to measure effects of 1,280 pharmacologically active compounds on complex phenotypes in isogenic cancer cell lines which harbor activating or inactivating mutations in key oncogenic signaling pathways. Using multiparametric chemical–genetic interaction analysis, we observed phenotypic gene–drug interactions for more than 193 compounds, with many affecting phenotypes other than cell growth. We created a resource termed the Pharmacogenetic Phenome Compendium (PGPC), which enables exploration of drug mode of action, detection of potential off-target effects, and the generation of hypotheses on drug combinations and synergism. For example, we demonstrate that MEK inhibitors amplify the viability effect of the clinically used anti-alcoholism drug disulfiram and show that the EGFR inhibitor tyrphostin AG555 has off-target activity on the proteasome. Taken together, this study demonstrates how combining multiparametric phenotyping in different genetic backgrounds can be used to predict additional mechanisms of action and to reposition clinically used drugs.**