npj | Systems Biology and Applications

www.nature.com/npjsba

ARTICLE OPEN

L1000CDS^2: LINCS L1000 characteristic direction signatures search engine

Qiaonan Duan[1,2,5], St Patrick Reid[3,5], Neil R Clark[1,2], Zichen Wang[1,2], Nicolas F Fernandez[1,2], Andrew D Rouillard[1,2], Ben Readhead[2], Sarah R Tritsch[3], Rachel Hodos[2], Marc Hafner[4], Mario Niepel[4], Peter K Sorger[4], Joel T Dudley[2], Sina Bavari[3], Rekha G Panchal[3] and Avi Ma'ayan[1,2]

The library of integrated network-based cellular signatures (LINCS) L1000 data set currently comprises of over a million gene expression profiles of chemically perturbed human cell lines. Through unique several intrinsic and extrinsic benchmarking schemes, we demonstrate that processing the L1000 data with the characteristic direction (CD) method significantly improves signal to noise compared with the MODZ method currently used to compute L1000 signatures. The CD processed L1000 signatures are served through a state-of-the-art web-based search engine application called L1000CDS^2. The L1000CDS^2 search engine provides prioritization of thousands of small-molecule signatures, and their pairwise combinations, predicted to either mimic or reverse an input gene expression signature using two methods. The L1000CDS^2 search engine also predicts drug targets for all the small molecules profiled by the L1000 assay that we processed. Targets are predicted by computing the cosine similarity between the L1000 small-molecule signatures and a large collection of signatures extracted from the gene expression omnibus (GEO) for single-gene perturbations in mammalian cells. We applied L1000CDS^2 to prioritize small molecules that are predicted to reverse expression in 670 disease signatures also extracted from GEO, and prioritized small molecules that can mimic expression of 22 endogenous ligand signatures profiled by the L1000 assay. As a case study, to further demonstrate the utility of L1000CDS^2, we collected expression signatures from human cells infected with Ebola virus at 30, 60 and 120 min. Querying these signatures with L1000CDS^2 we identified kenpaullone, a GSK3B/CDK2 inhibitor that we show, in subsequent experiments, has a dose-dependent efficacy in inhibiting Ebola infection *in vitro* without causing cellular toxicity in human cell lines. In summary, the L1000CDS^2 tool can be applied in many biological and biomedical settings, while improving the extraction of knowledge from the LINCS L1000 resource.

npj Systems Biology and Applications (2016) **2,** 16015; doi:10.1038/npjsba.2016.15; published online 4 August 2016

INTRODUCTION

Systematic collection and analysis of genome-wide gene expression drug–response data from human cell lines can be used to identify drug repurposing opportunities, discover novel mechanisms of action for compounds, realize small-molecule mimickers of endogenous ligands, and assist in predicting side effects for pre-clinical compounds.[1] Such an approach was initially made possible by the Connectivity Map,[2] which contains the data on the transcriptional responses of four human cancer cells lines 6 h after exposure to one of ~1,300 drugs and other small molecules. The Connectivity Map project has been extended under the auspices of the NIH library of integrated network-based cellular signatures (LINCS) program[3] by using a cost-effective genome-wide transcriptomics assay based on Luminex bead technology called L1000. Upon the completion of Phase I of the LINCS program, LINCS-L1000 data are available on the responses of ~50 human cell lines to one of ~20,000 compounds across a range of concentrations for a total of over one million experiments.

The process of computationally extracting signatures from mesenger RNA expression data, such as the LINCS-L1000 data, can be accomplished using a variety of statistical methods. Currently, signatures from the LINCS-L1000 data are computed using the moderated *Z*-score (MODZ) method. Recently, we developed a multivariate method to compute signatures called the Characteristic Direction (CD).[4] The CD method gives less weight to individual genes that display a large change in magnitude when comparing two conditions, for example, comparing gene expression from drug-treated cells with control cells. Some genes that change in magnitude substantially may be given a lower score, or a *P* value, compared with other methods such as the fold-change method. In fact, the fold-change method only considers the change in magnitude; and it is known to perform poorly. The CD method gives more weight to genes that move together in the same direction across repeats. So, a gene that changes less but 'moves' together with a large group of other genes may be scored higher than a gene that changed more in overall magnitude. The method first identifies the linear hyperplane that best separates the control samples from the treatment samples using linear discriminant analysis, and then uses the Normal to this hyperplane to define the direction of change in expression space for each gene. We have shown before that the CD method is more

[1]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [2]Department of Genetics and Genomics Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [3]US Army Medical Research Institute of Infectious Diseases, Frederick, MD, USA and [4]Department of Systems Biology, Harvard Medical School, Boston, MA, USA.
Correspondence: A Ma'ayan (avi.maayan@mssm.edu)
[5]These authors contributed equally to this work.
Received 6 October 2015; revised 3 April 2016; accepted 5 May 2016

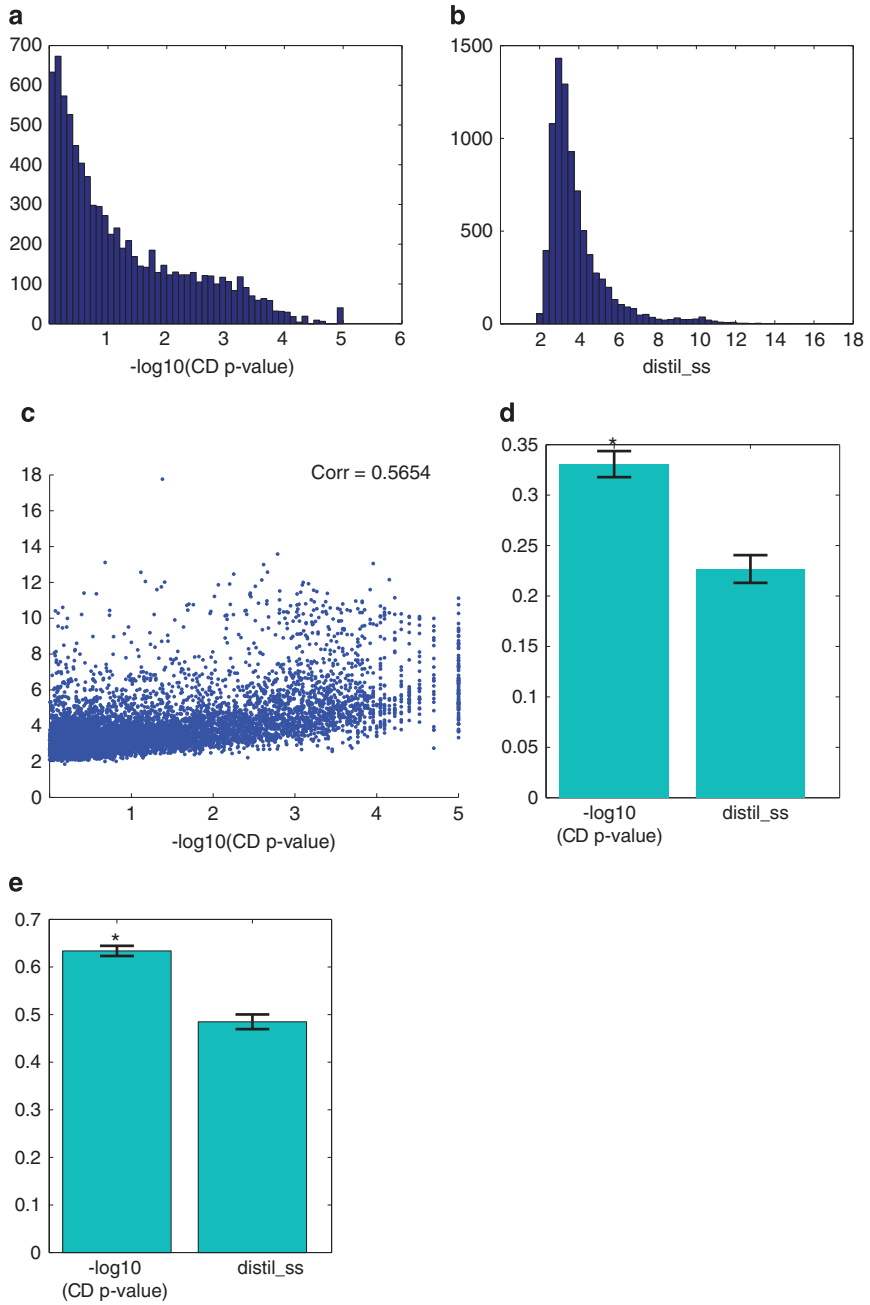


Figure 1. Intrinsic benchmarking. Expression signatures for each small molecule are computed with the Characteristic Direction (CD) algorithm or downloaded from lincscloud.org. The signatures on lincscloud are computed using the Moderated *Z*-score (MODZ) method. (**a**, **b**) Histograms of the significance scores for the 8,301 signatures from the LJP5 and LJP6 batches. (**c**) Correlation between the strength metrics for signatures computed by the two methods. (**d**) Correlation between differential expression significance rank and dose rank using the two methods of computing differential expression. (**e**) Correlation between differential expression significance rank and dose rank using the two methods of computing differential expression without the influence of insignificant perturbations.

sensitive in identifying the 'correct' differentially expressed genes than most popular alternative methods using several benchmarking strategies applied to the real data. With these benchmarks the CD method outperformed limma,[5] DESeq,[6] significant analysis of microarrays[7] and the *t*-test.

In this study, we applied the CD method to process the LINCS-L1000 data. We demonstrate that with the CD method we can significantly extract better signatures compared with the currently available computed signatures. All of the benchmarks presented in this current manuscript are different from the benchmarks presented before. These new benchmarks are specific for the L1000 data, and as such, they set the stage for other methods to be developed by showing how internal and external data can be used to evaluate computational pipelines for processing the L1000 data. To enable access and utility to the composite of the reprocessed CD signatures, we developed a state-of-the-art web-based application, which is a signature-search-engine called L1000CDS^2. L1000CDS^2 computes the angle between an input signature vector and the LINCS-L1000 data to prioritize small molecules and drugs to either reverse or mimic observed changes in gene expression. With the L1000CDS^2 tool, we prioritized small molecules that can potentially reverse gene expression in 670

disease signatures extracted from gene expression omnibus (GEO), and predicted small molecules as potential mimickers of endogenous ligands. Using an independent collection of gene expression signatures of single-gene perturbations extracted from GEO we also predicted the most likely targets for each small molecule profiled by the L1000 assay. As a case study, we tested one prediction experimentally. Gene expression profiles from human dendritic cells soon after infection with Ebola virus (EBOV) were submitted as input to L1000CDS[2]. The top candidate from the L1000CDS[2] signature search was kenpaullone, a non-specific inhibitor of kinases including GS3Kβ and CDK2.[8] Subsequent experiments on EBOV-infected HeLa and human foreskin fibroblast cells demonstrated dose-dependent inhibition by kenpaullone. Gene-set enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG)[9] pathways, the gene ontology[10] biological process tree, the mouse genome informatics (MGI) mammalian phenotype ontology, ChIP-x enrichment analysis (ChEA),[11] the Encyclopedia of DNA Elements (ENCODE),[12] kinase enrichment analysis (KEA)[13] and Expression2Kinases[14] confirms the potential involvement of GSK3β and CDK2 in early EBOV infection and suggest that kenpaullone induces innate immune response genes to assist infected cells to detect the virus. These results suggest new means to combat EBOV and potentially other pathogen infections, and demonstrate the utility of L1000CDS[2].

RESULTS

Benchmarking the characteristic direction method

As we previously showed that the CD method significantly improves the quality of extracted gene expression signatures from published microarray and RNA-seq studies,[4] we hypothesized that this method would significantly improve the extraction of signatures from the L1000 data set. To test this hypothesis, we benchmarked L1000 signatures extracted with the CD method with the only currently implemented method called moderated *Z*-scores MODZ. First, we examined the overall distribution of CD signature *P* values and MODZ signature strengths (distil_ss) for a collection of 8,301 unique L1000 experiments from two batches (LJP5 and LJP6; Figure 1a,b). These histograms show a much fatter tail for the signatures extracted with the CD method, suggesting that this method identifies more potentially significant signatures. Indeed, there are 685 significant signatures that pass the distil_ss >6 cutoff, which is currently recommended to call significant MODZ signatures, compared with 2,045 significant signatures called by the CD method with a $P<0.01$. As there is a one-to-one correspondence between the signatures computed by the two methods, we computed the correlation between the two scoring schemes. We observed mild but significant correlation between the two signature-scoring methods (Figure 1c). This suggests that there is some overlap between the methods but also there are significant differences.

Out of the 8,301 experiments in LJP5 and LJP6, there are 1,415 unique perturbation conditions if dose is not considered. We examined the overall correlation between dose and overall change computed by the two methods. The results from this analysis show that the CD method identifies higher correlation between dose and overall change compared with the MODZ method (*P* value = 3.88e − 08, *t*-test; Figure 1d). This correlation improves, and the difference in performance between the two methods widens, when filtering out all the non-significant perturbations (*P* value = 1.1782e − 14, *t*-test; Figure 1e). For the next benchmarking assessment we analyzed the data from the LJP4 batch where four classes of endogenous ligands: growth factors; cytokines; interferons and others, were systematically applied to six breast cancer cell lines. We asked which signature extraction method separates the signatures by perturbation type, time point and cell line. Using two unsupervised clustering methods: multidimensional scaling and hierarchical clustering, we plotted the signatures while coloring each signature by its known class membership: perturbation type, time point and cell line (Supplementary Figures S1 and S2). These figures show that the clustering results computed with the CD better reflect known biological knowledge about the experimental conditions applied. The above benchmarking methods rely only on the L1000 data and some limited knowledge about the perturbation and this is why we call this set of benchmarks intrinsic. Additional benchmarking methods can be achieved with external data, termed extrinsic benchmarks. For this, we next asked whether similar signatures induced by different small molecules also share chemical structural similarity.

Experiments selected for this benchmarking were performed on nine core cell lines: HA1E, VCAP, HCC515, PC3, A375, HEPG2, HT29, MCF7, A549, upon treatment with 10 μM and where gene expression was measured at 6 or 24 h covering 20,412 small-molecule compounds. For experiments of identical conditions, only the strongest as measured by CD signature *P* value were kept. To compute the similarity between gene expression signatures of these experiments, the cosine distance was used to for both CD and MODZ signatures. Euclidean distance was also used as another measure of similarity between MODZ signatures. The chemical structure of the 20,412 small molecules were encoded into the 166-bit Molecular ACCess System (MACCS) fingerprint, or the Extended-Connectivity Fingerprints (ECFP4). Tanimoto score was used to quantify fingerprint similarity with Tanimoto coefficient of larger than 0.9 as a cutoff. Although there is some relationship between expression signatures computed with the MODZ method and the similarity of the chemical structure of the perturbagens, the CD method is able to recover more significant correlations between structure and expression (Figure 2a). The benchmarking strategy also highlights that the ECFP4 molecular fingerprint method outperforms the MACCS method.

The next extrinsic benchmarking approach asks whether direct protein–protein interactions (PPI) of known drug targets are also differentially expressed after drug treatment. A recent study that utilized the original Connectivity Map data set showed how the protein interactions of known targets are commonly found in the differentially expressed genes induced by a drug treatment.[15] CD signatures were compared with MODZ signatures for their ability to prioritize differentially expressed genes that are also direct protein interactors of the drug targets. The signatures of 756 small molecules in the LINCS L1000 data that have at least one known protein target were selected for this benchmarking strategy. Differentially expressed genes computed by each method, CD or MODZ, were sorted by their absolute differential expression value and compared with direct PPI of the targets using a random walk (Figure 2b). The aggregated walks for the CD signatures show a much sharper peak compared with the signatures computed with MODZ, suggesting that the CD detects more direct protein–protein interactions of the known targets from the L1000 LINCS data. In summary, the intrinsic and extrinsic benchmarking analyses clearly demonstrate that the CD is potentially a better alternative for computing signatures compared with the MODZ method. However, this result is global. The CD method works better than the MODZ method across many data sets in general, but it is possible that the MODZ method consistently outperforms the CD under some specific conditions. In addition, caution should be placed when considering individual genes when using the CD method, for example, when picking genes for RT-PCR validation. This is because the CD method can score highly genes that will display changes that would not be considered statistically significant with univariate methods such as the student's *t*-test. Although it appears that the CD performs better in ranking the differentially expressed genes and determining which signatures should be considered significant, more benchmarks are needed.

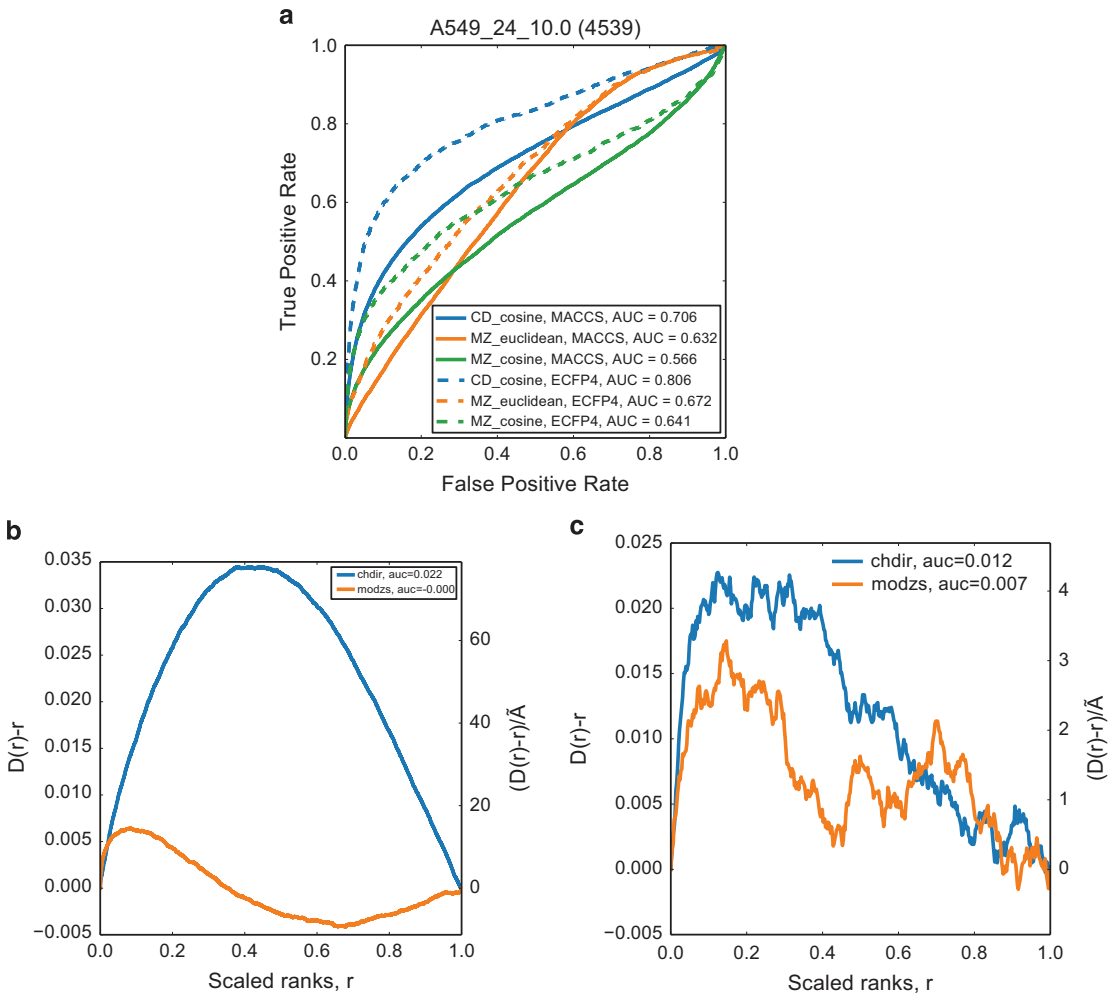


Figure 2. Extrinsic benchmarking. (**a**) ROC curves showing the recovery of structurally similar small-molecule compounds compared with gene expression signature similarities in A549 cells after 24 h treatment with 10 μM of all compounds computed using the two different methods: the cosine distance between Characteristic Direction (CD) signatures in blue, and the Euclidean distance of the Modulated Z-score (MODZ) signatures in orange, and cosine distance of MODZ signatures in green. Chemical fingerprints similarities used to benchmark the gene expression signature similarity are MACCS and ECFP4, plotted in solid and dashed curves, respectively. (**b**) The deviation from the cumulative distribution of a uniform for the rankings of drug targets and their direct interactors in gene expression signatures computed using CD (blue) and MODZ (orange) under the same conditions. (**c**) Recovery of known drug targets by observing the ranks of gene expression signatures extracted from GEO ($n = 2206$) where 917 genes were perturbed by either knocked-down, knocked out, or over-expressed in mammalian cells. GEO signatures are ranked by cosine distance when queried with the L1000 LINCS data processed by the MODZ or the CD methods. The deviation from the cumulative distribution of a uniform for the rankings of drug targets as determined by DrugBank where signatures are computed using CD (blue) and MODZ (orange) under the same conditions. ECFP4, extended-connectivity fingerprints; MACCS, molecular access system; ROC, reveiver operating curves.

For the third extrinsic benchmark we examined which method directly recovers the known drug targets using an independent data set of gene expression signatures extracted from the GEO database. We collected 2,206 signatures for 917 unique genes from studies where a single gene was perturbed (knocked-down, knocked out or over-expressed) in mammalian cells. Using the cosine distance, we ranked the GEO signatures for their similarity with the L1000 LINCS data processed by the MODZ or the CD method and evaluated where the known targets, as listed in DrugBank, are ranked. We see that the CD method ranks known targets higher than the MODZ method (Figure 2c). Hence, this benchmark also supports that the CD method performs better than the MODZ. As knowing the potential targets of small molecules is very useful for many drug discovery applications, the predicted targets for all small-molecule signatures, using the CD method, were added as predictions to the L1000CDS^2 web-based search engine software application.

The L1000CDS^2 web-based search engine software application

The input page on the L1000CDS^2 web application consists of five sections: input textboxes; example signatures; configuration; metadata and recent searches (Figure 3). The entry point into L1000CDS^2 is to paste up/down gene lists into the up/down gene text boxes, or paste a signature into the up gene text box. A signature is a list of genes and their differential expression values separated by a comma. The search button will only become enabled when both the up/down gene-set text boxes are filled, or when the up gene-set text box is filled with a signature. Clicking the Search button will return the top 50 signatures in a table on the results page. The example signatures section includes pre-computed signatures that users can submit as input. The EBOV signatures are the gene expression signatures described in this pager. The disease signatures comprise of 670 disease signatures extracted manually from GEO by identifying the control and disease sample GSM files. The ligand signatures are consensus

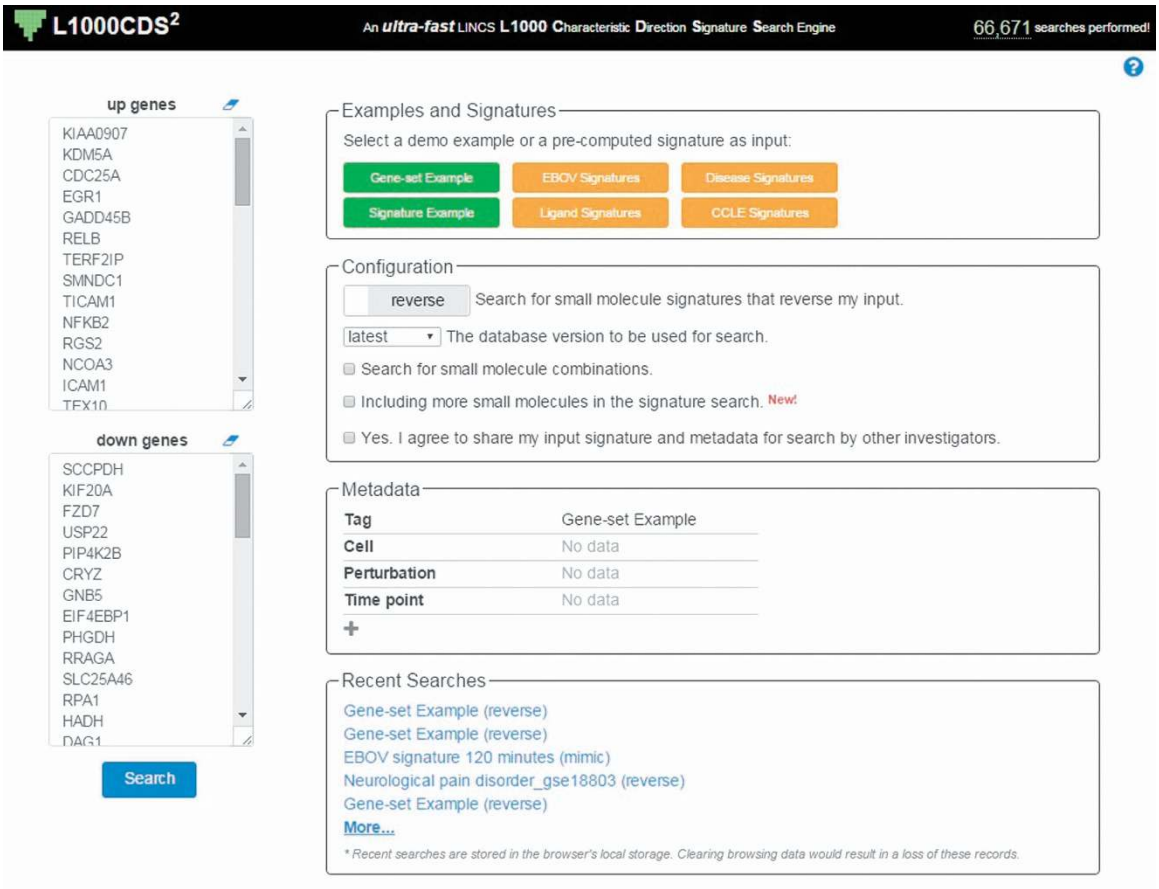
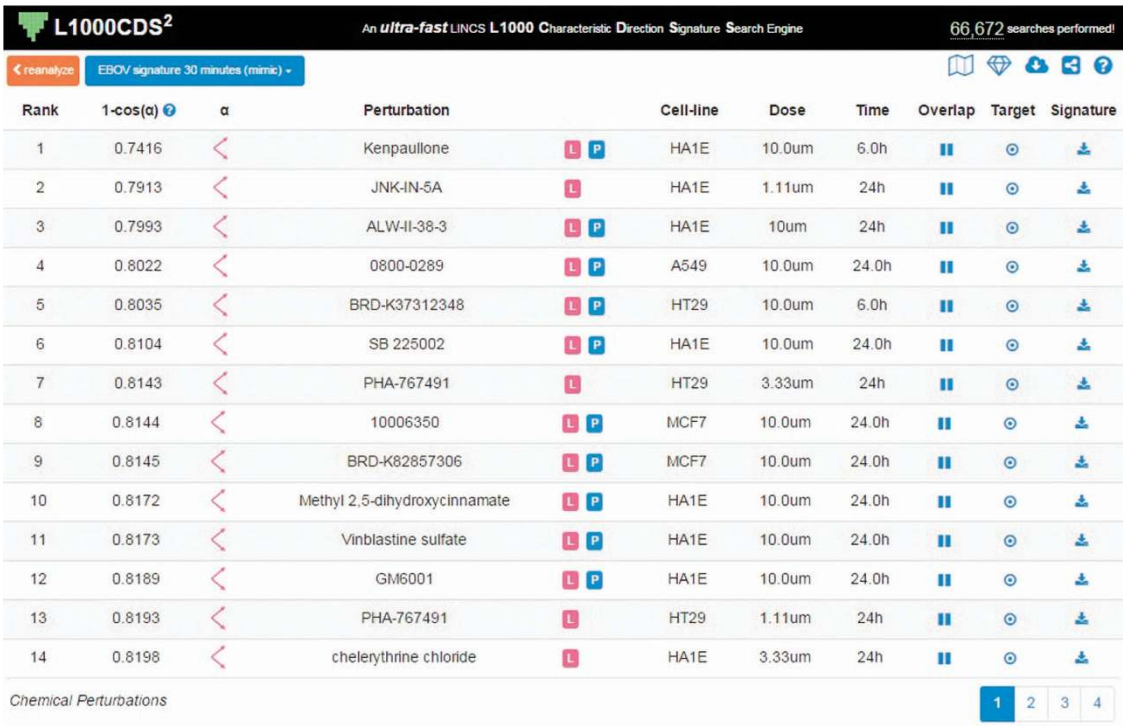
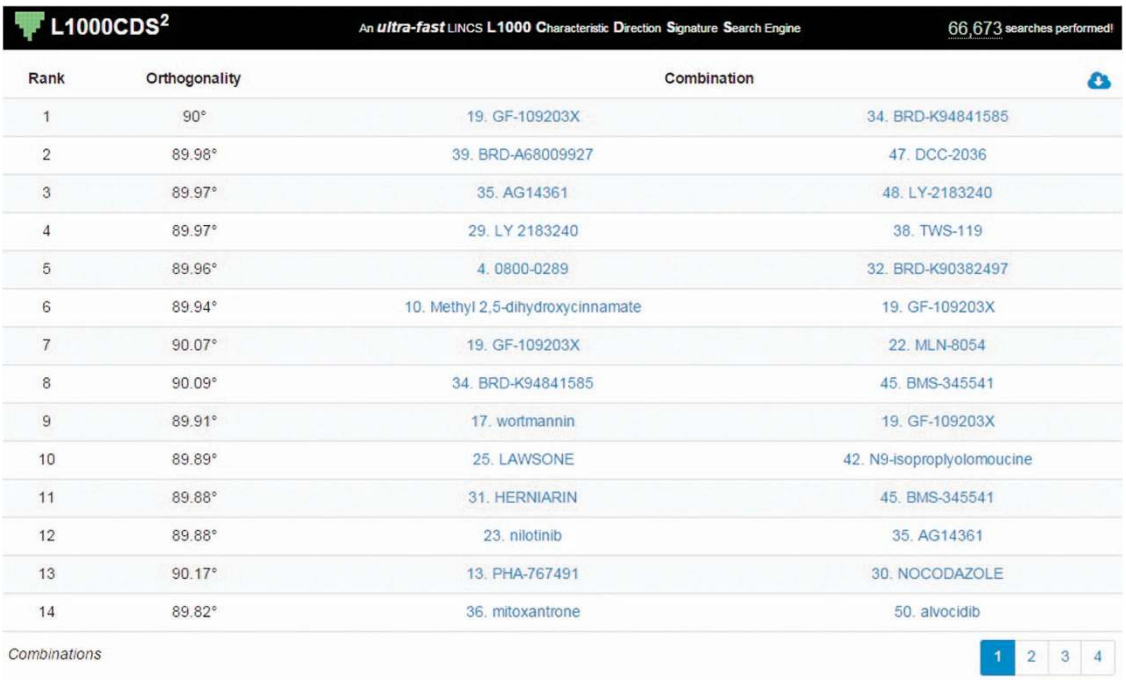


Figure 3. Screenshot from the input page of the L1000CDS^2 software application. The input text boxes toggle between up and down sets, or an input vector option. Canned analysis for 670 disease signatures is provided with few clicks. The Ebola, ligand and cancer cell line signatures are provided as canned examples.



Rank	1-cos(α)	α	Perturbation		Cell-line	Dose	Time	Overlap	Target	Signature
1	0.7416		Kenpaullone	L P	HA1E	10.0um	6.0h			
2	0.7913		JNK-IN-5A	L	HA1E	1.11um	24h			
3	0.7993		ALW-II-38-3	L P	HA1E	10um	24h			
4	0.8022		0800-0289	L P	A549	10.0um	24.0h			
5	0.8035		BRD-K37312348	L P	HT29	10.0um	6.0h			
6	0.8104		SB 225002	L P	HA1E	10.0um	24.0h			
7	0.8143		PHA-767491	L	HT29	3.33um	24h			
8	0.8144		10006350	L P	MCF7	10.0um	24.0h			
9	0.8145		BRD-K82857306	L P	MCF7	10.0um	24.0h			
10	0.8172		Methyl 2,5-dihydroxycinnamate	L P	HA1E	10.0um	24.0h			
11	0.8173		Vinblastine sulfate	L P	HA1E	10.0um	24.0h			
12	0.8189		GM6001	L P	HA1E	10.0um	24.0h			
13	0.8193		PHA-767491	L	HT29	1.11um	24h			
14	0.8198		chelerythrine chloride	L	HA1E	3.33um	24h			

Figure 4. Screenshot from the single drug/small-molecule results page of the L1000CDS^2 software application.



Rank	Orthogonality	Combination	
1	90°	19. GF-109203X	34. BRD-K94841585
2	89.98°	39. BRD-A68009927	47. DCC-2036
3	89.97°	35. AG14361	48. LY-2183240
4	89.97°	29. LY 2183240	38. TWS-119
5	89.96°	4. 0800-0289	32. BRD-K90382497
6	89.94°	10. Methyl 2,5-dihydroxycinnamate	19. GF-109203X
7	90.07°	19. GF-109203X	22. MLN-8054
8	90.09°	34. BRD-K94841585	45. BMS-345541
9	89.91°	17. wortmannin	19. GF-109203X
10	89.89°	25. LAWSONE	42. N9-isoproplyolomoucine
11	89.88°	31. HERNIARIN	45. BMS-345541
12	89.88°	23. nilotinib	35. AG14361
13	90.17°	13. PHA-767491	30. NOCODAZOLE
14	89.82°	36. mitoxantrone	50. alvocidib

Figure 5. Screenshot from the drug pair results page of the L1000CDS^2 software application.

ligand perturbations calculated from the LINCS L1000 LJP4 subset. The cancer cell line encyclopedia (CCLE)[16] signatures were computed from the CCLE gene expression data by comparing the gene expression profile of each cell line to the rest using the CD method. The configuration section provides several options to customize a search. The mimic/reverse slider chooses between searching for small molecules that mimic the input signature or reverse it. The default search mode is reverse. Small-molecule pairwise combinations search is also supported. Users can share their input signatures and metadata so others can query their signatures. In the metadata section, any metadata associated with the input signature can be entered. Most importantly users are encouraged to enter at least one tag for future reference. The 20 most-recent searches are stored in the recent search section. Clicking on an entry reloads recent search results.

On the result page, the search results are rendered as a paginated table with 14 entries per page (Figure 4). Each entry provides seven columns of information about the signature: rank; score; perturbation; cell-line; dose; time point and overlap with the input. For the gene-set search, the search score is the overlap between the input DE genes and the signature DE genes divided by the effective input. The effective input is the length of the intersection between the input genes and the L1000 genes, as some input lists contain genes that are not present in the L1000 data set. This includes all ~22,000 L1000 genes, not just the measured ~1,000.

Clicking the overlap button will show the overlapping genes (and their values) in two text boxes. If the user input-type is up/down gene lists, then the first box will show the overlapping genes between the input up genes and the signature up (down) genes, and the second will show the overlap between the input down and the signature down (up) in the mimic (reverse) mode. If the input is a signature, then the first box will show genes with a positive value from the input signature, and the second box the negative value genes. The signature values and input values in both boxes are expected to be mostly of the same sign in the mimic mode, and mostly the opposite sign in reverse mode. The Enrichr button under each text box will send the genes to Enrichr for enrichment analysis. Clicking the download button will download all the information about a signature as a JavaScript Object Notation (JSON) file.

On top of the table are buttons and icons that provide various useful services. The reanalyze button redirects the user back to the input page with the submitted lists or signatures preloaded in the input textboxes. Users can then reanalyze their input using different configurations, or modify the associated metadata. This function also has a bearing on sharing results with others. It provides a way for users to reanalyze their input with different settings and obtain a permanent URL for each analysis. The tag button displays the tag and search mode. Clicking on the button shows the input metadata. The cloud download icon downloads the results table as a .csv file. The diamond icon performs enrichment analysis on the substructures of the top ranked small molecules. The results of the substructure enrichment analysis are displayed as a table where each row is a significantly enriched substructure. Each row provides three pieces of information: substructure, *P* value and perturbation count. The substructure is represented as a string in the SMARTS format. The *P* value is computed using the Fisher's exact test. The perturbation count shows the number of perturbations that have the same substructures. Clicking on the share icon produces a permanent URL that can be used to share the substructure enrichment analysis results through an e-mail, a publication or other documentation. Clicking on the plus sign shows a visualization of the substructure and a table of the top perturbations that contain the substructure. The rank in the table is the same rank of the perturbations in the top 50 results table.

If the user chooses to search for small-molecule combinations, then a table of signature combinations will appear below the single perturbation result table (Figure 5). This table is also a paginated table with 14 entries per page. Each entry provides three pieces of information about the identified combinations: rank; synergy score and combinations. When searching for combinations, L1000CDS^2 compares every possible pair among the top 50 matching signatures and computes the potential synergy between each pair by examining the level of orthogonality. With the gene-set search, the synergy score is calculated as the combined overlap of the differentially expressed

genes of the two drug signatures with the input gene lists. In a cosine distance search, the synergy is calculated as the orthogonality between two CD signatures. The rational for this is that if two perturbations are orthogonal, then they may impart their overall effect through two independent pathways. The rank is based on the orthogonality score. The number before each chemical perturbation in the combinations column is the rank of that perturbation in the single signature result table. Clicking on a perturbation will highlight that perturbation in the single signature results table so the user can learn more details about that perturbation. Clicking on the cloud download button on the upper right corner downloads the combination table as a .csv file.

L1000CDS^2 predicts kenpaullone as a potential drug to inhibit Ebola infection

To further utilize the functionality of L1000CDS^2 we applied it to predict drugs and small molecules that can potentially inhibit Ebola infection. To accomplish this we first collected expression signatures from human cells infected with Ebola. Dendritic cells from four human donors were infected with Ebola, and then genome-wide gene expression was measured prior to infection and at 30, 60 and 120 min after infection. The differentially expression genes at each time point were computed comparing each time point to the control. This resulted in 1,031 significantly differentially expressed genes at 30 min, 746 at 60 min and 248 at 120 min. These signatures offer an opportunity to further investigate the early response to Ebola infection compared with most few other available previously published genome-wide expression data sets before and after Ebola infection in human and non-human primate cells.[17–21] We first analyzed the differentially expressed gene lists using Enrichr,[22] an online tool to perform enrichment analysis. We found that the upregulated differentially expressed genes are enriched in immune response terms, and hypothesized that small molecules that mimic this gene expression state may enhance the intrinsic immune response to assist in attenuating Ebola infection. We queried these signatures with the L1000CDS^2 tool to prioritize drugs and small molecules that can potentially mimic gene expression in Ebola-infected cells. Cosine distances were calculated for each L1000 direction with an Ebola signature direction from the three time points. Perturbations were then sorted by cosine distance, prioritizing similar signatures. Interestingly, kenpaullone was found to be ranked first for all three time points even though each time point had a unique set of differentially expressed gene signatures (Tables 1 and 2). It should be noted that using the existing signature search tool available on the new lincscloud Connectivity Map website, kenpaullone was ranked in the top 50 for the 60 and 120 time point signatures with ranks of 39th and 23th, respectively. These are relatively high ranks, but likely not high enough to prioritize kenpaullone as a top choice for experimental validation.

We next sought to assess whether several of the top ranked small molecules, including kenpaullone, can inhibit Ebola infection in tissue culture. For an initial screen we pretreated HeLa cells with 20 μM of each top ranked small molecule and then infected the cells with Ebola at a multiplicity of infection (MOI) of five for 48 h. Ebola-infected cells were stained for viral antigen and analyzed on an Opera, confocal high-content imaging platform. Under these conditions kenapaullone was observed to inhibit Ebola infection by ~52% (Figure 6a). The small-molecule SB218078, a Chk1 inhibitor, inhibited Ebola by 60%. However, follow-up studies using this small molecule yielded inconclusive results owing to issues of solubility (data not shown). The remaining small molecules, AG-14361, Menadione and Methyl 2, 5-didroxycinnamate had minimal inhibitory affect against the virus and under these conditions whereas daunorubcin was observed to be cytotoxic.

Table 1. Top five predicted drugs at each time point

Drug name	*Cosine distance*	*Batch*	*Cell line*	*Time point (h)*	*Concentration (μM)*
30 min					
Kenpaullone	1.2584	CPC002	HA1E	6	10
0800-0289	1.1978	CPC014	A549	24	10
BRD-K37312348	1.1965	CPC016	HT29	6	10
SB 225002	1.1896	CPC001	HA1E	24	10
10006350	1.1856	CPC012	MCF7	24	10
60 min					
Kenpaullone	1.2756	CPC002	HA1E	6	10
PD 166793	1.2619	CPC001	HA1E	24	10
BAPTA-AM	1.2564	CPC001	HA1E	24	10
BRD-U74615290	1.2396	CPC014	HT29	6	10
NCGC00229596-01	1.2303	CPC008	HT29	6	10
120 min					
Kenpaullone	1.3489	CPC002	HA1E	6	10
NSC 23766	1.3478	CPC006	PC3	24	160
Rosiglitazone	1.3386	CPC006	HA1E	24	80
7-hydroxy-2, 3, 4, 5-tetrahydro-1H-[1] benzofuro[2, 3–c] azepin-1-one	1.3351	CPC007	A549	24	10
LY 364947	1.3183	CPC003	PC3	24	10

Table 2. Top 10 predicted drugs by their rank product, i.e., multiplying the ranks across the three time points as determined by the cosine distance

Broad molecule ID	*Drug*	*Rank product*
BRD-K37312348	Kenpaullone	1
BRD-K40919711	BAPTA-AM	252
BRD-A97437073	Rosiglitazone	1,755
BRD-A68009927	Daunorubicin hydrochloride	2,880
BRD-A06352508	SB 218078	3,136
BRD-K78126613	MENADIONE	3,328
BRD-K88741031	Methyl 2, 5-dihydroxycinnamate	4,284
BRD-K00615600	AG14361	5,566
BRD-K31342827	GF-109203X	8,385
BRD-A75409952	wortmannin	17,204

Taken together, out of the five small molecules selected for the initial screen, kenapaullone remained to be the most promising.

To determine whether kenpaullone can inhibit Ebola infection in a dose-dependent manner, we conducted a series of dose–response experiments. HeLa cells were pretreated with kenpaullone with doses ranging from 0.3 to 75 μM and then infected with Ebola at an MOI of five for 48 h. In these experiments kenpaullone was observed to consistently inhibit Ebola infection in a dose-dependent manner (Figure 6b,c). It is important to note that while the compound potently inhibited viral infection, it did not affect overall cell number and was therefore non-cytotoxic. Since kenpaullone is a known GSK3B inhibitor, we next experimentally tested other well-established GSK3 inhibitors for efficacy of inhibiting Ebola infection in HeLa cells. These compounds were SB216763, SB415286, TCS2002 and TC-G 24. The cells were pretreated with the drug for 2 h as before, and then infected with Ebola at the same MOI of five. By cell number, the compounds were non-toxic (only SB216763 showed minimal toxicity at 100 μM), but these compounds had minimal effect on blocking or enhancing Ebola infection. These results suggest that kenpaullone likely exerts its effects not only by inhibiting GSK3B but also through CDK2 and likely other

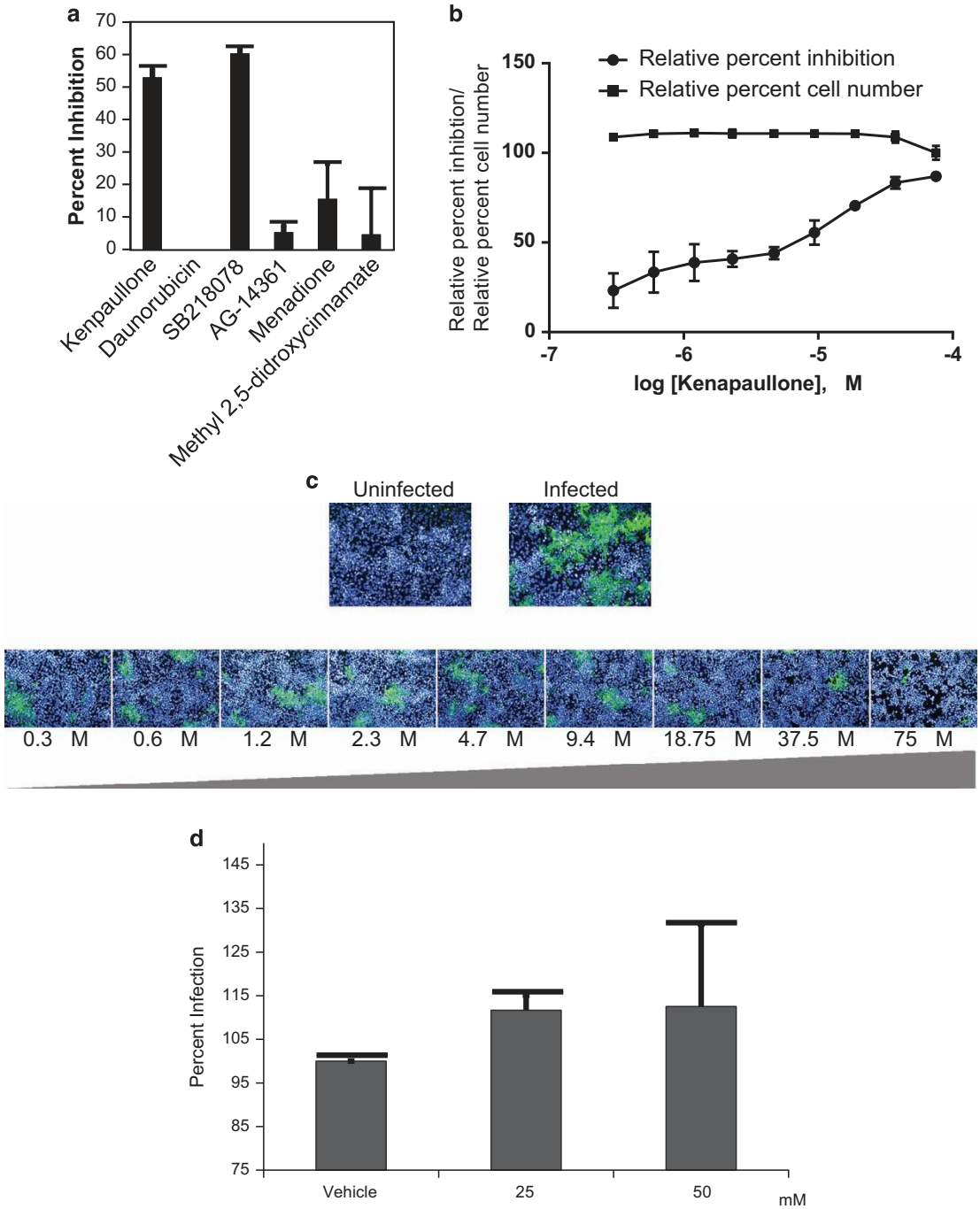


Figure 6. Experimental validation of small-molecule predictions. (**a**) Initial screen of the top five predicted small molecules to attenuate Ebola infection. HeLa cells were treated with 20 μM of each small molecule and then infected with Ebola at a multiplicity of infection (MOI) of five for 48 h. Ebola-infected cells were stained for viral antigen and analyzed on a confocal high-content imaging platform. (**b**) Dose–response experiments. HeLa cells were pretreated with a dose range of kenpaullone (0.3–75 μM) then infected with Ebola at an MOI of five for 48 h. (**c**) Representative images of cells treated in **b**. (**d**) Pre-treatment of HeLa cells with NCGC00184902-01 at two doses infected with Ebola. NCGC00184902-01 was predicted to reverse expression of the Ebola infection signatures at all three time points.

mechanisms. We also experimentally tested several small molecules that were consistently predicted to reverse expression across all three time points. These compounds included: PAC-1, Honokiol and NCGC00184902-01. Among these, NCGC00184902-01, showed slight enhancement of Ebola infection while the others two compounds had no effect (Figure 6d). Finally, we tested whether the effect of kenpaullone is cell-type specific by testing its efficacy to attenuate Ebola infection in another cell types. Human foreskin fibroblasts, considered a primary human cell line, were pretreated with kenpaullone and infected with Ebola exactly in the same way as was done for HeLa cells. We observed a dose-dependent inhibition of infection in human foreskin fibroblasts as well (data not shown).

To identify potential molecular mechanisms induced by Ebola infection and mitigated by kenpaullone, we performed gene-set enrichment as well as Expression2Kinases (X2K) analyses. We used as input the up and down differentially expressed genes at 30, 60 and 120 min by Ebola, and the matching L1000 signatures

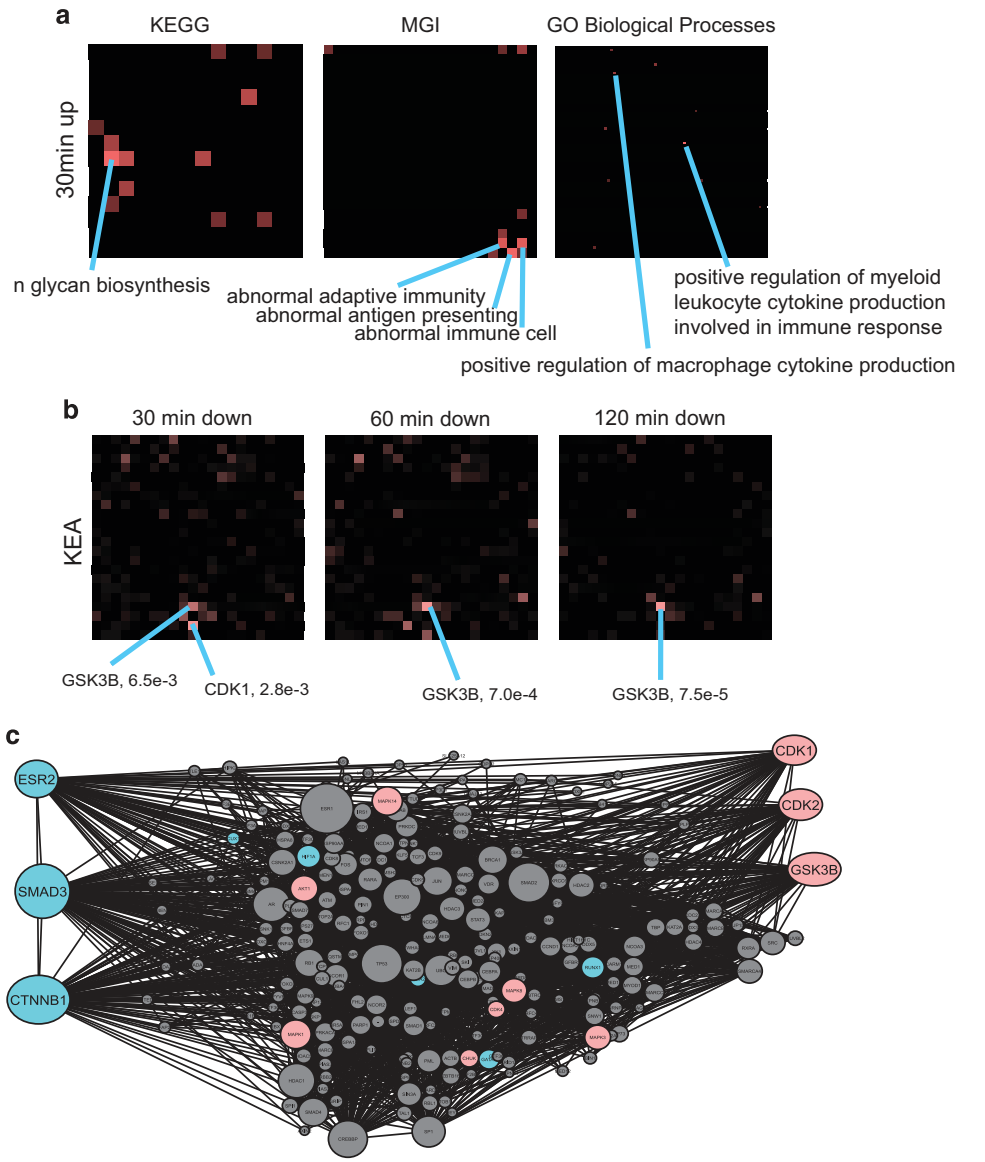


Figure 7. GO, KEGG, MGI, KEA and X2K enrichment analyses. (**a**) Gene Ontology, KEGG pathways and mammalian phenotype enrichment analyses visualization on three representative canvases for the upregulated genes after Ebola infection at 30 min. Each tile in each canvas represents a term/gene-set and where all terms are arranged based on their gene-set content similarity. The canvas is continuous so the sides fold on each other. The tiles brightness represents high enrichment scores (or low *P* values) computed with the Fisher's exact test. The most top enriched terms are highlighted. Complete results can be seen in supporting Table 1. (**b**) Kinase enrichment analysis visualized on a canvas where each tile represents a mammalian kinase and the gene sets for each kinase are its known substrates. The brightness of the tiles represent the enrichment *P* value scores computed using the Fisher's exact test. (**c**) Expression2Kinases analysis of the upregulated genes after 2 h. In this analysis, we first identify transcription factors that are enriched for targets within the differentially expressed genes based on prior ChIP-seq experiments. Then, the top ten transcription factors are connected through known protein–protein interactions. Finally, the resultant proteins within this subnetwork are subjected to kinase enrichment analysis with KEA. Node size reflects connectivity and color distinguishes transcription factors in blue, intermediate proteins in gray and kinases in green. GO, gene ontology.

with kenpaullone. These gene sets were submitted for enrichment and network analysis with the tools Enrichr[22] and Expression2Kinases.[14] The first set of enrichment analyses was applied to the upregulated genes induced by Ebola using the gene ontology biological process, KEGG pathways and the knockout mice mammalian phenotype ontology from the mouse genome informatics (MGI) gene-set libraries (Figure 7a, Supplementary Table S1). We observed a common theme. The upregulated genes across all three time points are associated with terms and pathways involving immune responses and *N*-linked glycosylation. For example, abnormal innate immunity is by far the most enriched MGI mouse phenotype term for the upregulated genes upon Ebola infection after 60 min (adjusted *P* value of < 0.0001, Fisher's exact test); whereas the most enriched KEGG pathway is the N-glycan biosynthesis (adjusted *P*-value of < 0.00001, Fisher's exact test). Next, we performed KEA[13] on the downregulated genes after Ebola infection. KEA includes kinase–substrate interactions extracted from publications and consolidated from six kinase–substrate open online databases. In this gene-set library each term is a kinase and its substrates are the gene sets associated with each term. Interestingly, applying KEA to the downregulated genes at all three time points shows most enrichment for known GSK3B substrates, the known target of kenpaullone (Figure 7b). At the 30-min time point, substrates of

CDK1 and CDK2 are also highly enriched together with GSK3B. These observation are consistent with our prediction of kenpaullone as the top mimicking drug. X2K analysis also confirms the potential involvement of GSK3B, CDK1 and CDK2 through the transcription factors beta-catenin, Smad3 and estrogen receptor 2 (Figure 7c). Smad3 activity was recently shown to be induced by kenpaullone to promote iTreg differentiation.[23] X2K analysis first identifies the enriched transcription factors upstream of the differentially expressed genes, then connects these transcription factors using known PPI collected from 20 databases, and then X2K applies KEA to the network of connected transcription factors. Hence, X2K attempts to connect changes in expression to their upstream regulatory mechanisms and cell signaling pathways through transcription factors and protein interactions. The results from the X2K analysis are consistent and robust to different thresholds and enrichment tests. Finally, we asked about the overlapping genes induced or repressed by both kenpaullone and early Ebola infection (Supplementary Table S2). The most enriched gene ontology terms, MGI phenotypes and KEGG pathways that are upregulated by both Ebola infection and kenpaullone are related to the innate immune response and include the following genes: *CIITA*, *LILRB1*, *RIPK1*, *CARD9*, *ALOX5AP*, *AHR*, *TLR4* and *CTSC*. These genes were reported to be regulated by STAT3 before based on two independent published ChIP-seq experiments from ENCODE[12] and ChEA.[11] STAT3 is a known inducer of the innate immune response[24] and we predict that STAT3 activity is further induced by kenpaullone. However, validation of such prediction remains for future studies.

DISCUSSION

Overall, we obtained initial coherent mechanistic insights about the potential intracellular pathways that may be required for early Ebola infection, and identified a small molecule that can potentially interfere with this process. Our finding that a GSK3B inhibitor can potentially attenuate Ebola infection is supported by some additional evidence in the literature. Inhibitors of GSK3B were previously reported to decrease the acute inflammatory response following sepsis, protect multiple organ injury and reduce muscle protein degradation during sepsis.[25,26] The glycolytic pathway implicated from our enrichment analysis is known to be linked to GSK3B activity and there is increasing appreciation that the glycolytic pathway is critical to the Ebola life cycle.[27] The non-specificity of kenpaullone makes studying its molecular mechanisms and direct targets difficult. However, here we observe that kenpaullone induces the expression of immune response genes and as such it is potentially a general anti-viral candidate. A relatively old study showed evidence that treatment with interferon can improve outcome and survival of non-human primates infected with Ebola.[28] However, interferons are known to cause severe side effects and may not be feasible to produce and deliver, that is why small molecules such as kenpaullone, which can potentially mimic/induce interferon are desired. The increase in the expression of innate immunity response genes by kenpaullone is predicted to assist infected cells and cells nearby infected cells to better sense the viral infection and respond quicker. It is still an open question whether kenpaullone interfere with the virus life cycle or just turns on the immune response. It is also not clear whether the induction of the viral innate immune response by kenpaullone is through the same pathway induced by Ebola or through a different mechanism. More specific small-molecule derivatives of kenpaullone have been developed,[29] and could be tested in future studies for potency for inhibiting Ebola, and other similar pathogens.

In summary, here we introduced an improved computational method that potentially elevates the usefulness of a subset of the newly generated publicly available LINCS-L1000 data set to rapidly prioritize small molecules that could reverse or mimic expression in disease and other biological settings. The large collection of signatures computed by the CD method is delivered as a state-of-the-art web-based application that is already widely used. Besides prioritizing small molecules for reversing or mimicking an input signature, or pre-computed signatures for 670 diseases and a collection of endogenous ligands, the L1000CDS^2 search engine web-based tool also predicts pairwise small-molecule combinations, performs substructure enrichment analysis and computes predicted targets based on an external set of signatures. The top predicted molecule to reverse expression in human cells infected with Ebola, kenpaullone, was shown to attenuate infection in a dose-dependent manner while not causing cellular toxicity in two cell lines. We also predicted the affected target genes and cell signaling pathways pointing to immune response genes driven by the inhibition of the CDK1-2 and GSK3B pathways, and potentially activating STAT signaling. However, kenpaullone should be tested in more cellular contexts before moving to animal models of Ebola. As kenpaullone is shown to induce an immune response that assists human cells to potentially combat Ebola infection, it is possible that this small molecule will have positive effects in blocking the spread of other viruses. The kenpaullone example is just the tip of the iceberg of what is possible with the LINCS L1000 data and the L1000CDS^2 web-based tool as many other applications are awaiting discovery.

MATERIALS AND METHODS

Computing LINCS L1000 characteristic direction signatures

LINCS L1000 level 3 normalized data, and level 5 moderated *Z*-scores (MODZ) data were downloaded from lincscloud.org and GEO (GSE70138). The normalized data set was processed with MATLAB (Natick, MA, USA) using customized version of the CD method.[4] A CD unit vector was calculated for each experiment replicate in comparison with all the control replicates on the same plate. A CD signature was computed for each experimental condition by averaging the CDs across replicates. The mean of the pairwise cosine distances between the CDs across replicates was used as a test statistic to assess the significance of a CD signature. Specifically, the mean was compared with a null distribution constructed from random sampling of irrelevant CD replicates to compute a *P* value. The differentially expressed genes were calculated using the random product algorithm.[30] The CD signatures and associated metadata are stored in a MongoDB (New York, NY, USA) database and available for download from the L1000CDS^2 website.

Benchmarking characteristic direction signatures with moderated *z*-score signatures

Dose-significance correlation. The distribution of CD *P* values and the distribution of MODZ strengths (distil_ss) were plotted as histograms for a collection of 8,301 unique L1000 experiments from the LINCS L1000 LJP5 and LJP6 subsets. The correlation between the *P* values and strengths was computed and graphed in a scatter plot. Out of the 8,301 experiments, 1,415 unique conditions (if dose was not considered) were identified. The correlations between dose and *P* value, and between dose and MODZ strength, were calculated for each of these conditions.

Ligand classification. Two hundred sixty two of the most significant signatures were selected from the collection of 1,374 signatures created from cancer cell lines treated with ligands in the LINCS L1000 LJP4 batch. Significant signatures were determined by the CD signature significance. Pairwise cosine distances between CD signatures and pairwise Euclidean distances between MODZ signatures of these same significant signatures were computed and organized into two distance matrices. Multidimensional scaling was applied to each matrix to visualize the signatures in a scatter plot. On the plots each signature was colored by its known ligand class membership, perturbation type, time point or cell-line. Two sub-matrices including only the significant signatures using the MCF7 cell-line were extracted from the two distance matrices. Hierarchical clustering with different distance measures was applied to these sub-matrices to generate dendrograms. The leaves of the dendrograms were labeled by ligand name and colored by their known class membership, perturbation type or time point.

Chemical–substructure/expression–signature correlations. Experiments conducted on nine core cell lines (HA1E, VCAP, HCC515, PC3, A375, HEPG2, HT29, MCF7, A549) upon small-molecule treatment at 10 μM for 6 or 24 h were selected from the LINCS L1000 data. These experiments cover 20,412 small-molecule compounds. The experiments were further filtered by only retaining the most significant unique signatures (cell line, time point, small molecule and dosage) using the CD signature *P* value. To compute the similarity between gene expression signatures of these filtered experiments, the cosine distance was used for both CD and MODZ signatures. Euclidean distance is also used as another measure of similarity between MODZ signatures. The 2D chemical structures of the 20,412 small-molecule compounds encoded in simplified molecular-input line-entry system (SMILES) format were retrieved and converted to the 166-bit MACCS fingerprint, or the ECFP4. Tanimoto similarity was used to quantify fingerprint similarity with Tanimoto coefficient larger than 0.9 as a cutoff. The similarity among substructures was compared with the similarity between gene expression signatures using Pearson's correlation analysis.

Differential expression of drug targets' direct PPI. The chemical structure of small-molecule compounds within the LINCS L1000 data encoded in the SMILES format were compared with those in DrugBank v4.3.[31] Nine hundred twenty seven were identified as FDA-approved drugs where 765 of them had at least one known target. The direct known PPI for these targets were collected from a literature-based mammalian PPI network covering 50,478 PPIs between 9,384 proteins as described in ref. 32. The MODZ and the CD signatures using the 765 drugs with known targets were extracted from the LINCS L1000 data according the following criteria: (1) The MODZ signature is strong, reproducible and self-connected as indicated by the is_gold field in the MODZ metadata; and (2) The CD signature of an experiment must be significant with $P \leqslant 0.05$ computed as indicated above. The genes in each CD or MODZ signature were sorted by their absolute value of differential expression for each drug, and then these ranks were compared to the known direct PPI for the drug's targets using a random walk. Aggregate average walks were plotted on line charts.

Predicting drug targets using an independent data set from GEO. To benchmark the ability of the CD or MODZ methods to predict drug targets using the LINCS L1000 gene expression data we examined 451 single-gene perturbation gene expression signatures curated from GEO where the perturbed gene is a known target for at least one drug profiled by the L1000 assay. Signatures from the LINCS L1000 drug perturbation data were first grouped by combining cell line, dose, and time point. The strongest signature for a given drug was selected based on reproducibility (is_gold) and signature strength (distil_ss). To rank the GEO signatures using the signatures from the LINCS L1000 computed by CD or MODZ, we sorted the GEO gene expression signatures in decreasing order of the absolute value of cosine similarity between the LINCS L1000 signatures and GEO signatures in the space of all shared genes for each LINCS L1000 signature. The ranks of true drug targets from Drugbank v4.3[33] were recorded and scaled to the total number of single drug target perturbations from GEO. The cumulative distribution of the scaled ranks was plotted to assess matches between targets from the gene perturbation profiles from GEO and the drug induced signatures from LINCS L1000.

Developing the L1000CDS^2 web-based search engine software application

The significant CD signatures from the LINCS L1000 CPC, CPD and LJP5-9 batches as determined by their *P* values ($P < 0.1$) were selected to construct the back-end database for search. This collection includes a total of 33,197 signatures. These signatures are stored in a MongoDB database dedicated to the app. To deploy L1000CDS^2 we implemented a search function that is hosted on an R server using the R Rook package (Nashville, TN, USA). The back-end of L1000CDS^2 is implemented in Node.js. The Node.js server fetches metadata from a MongoDB database and the server communicates with the client through JSON strings. The front-end is implemented using AngularJS (Mountain View, CA, USA) and Bootstrap (San Francisco, CA, USA). After the front-end receives the JSON string, it renders the results as paginated tables. To implement the substructure enrichment analysis, a matrix was constructed where the rows are the 166 MACCS fingerprints and the columns are the small molecules. The visualization of the 166 fingerprints were downloaded as PNG files from SMARTSviewer.[34] An R function was written to perform the enrichment analysis. This R function takes the top small-molecule IDs as input and performs the Fisher's exact test to compute enrichment scores.

Identifying kenpaullone as a potential drug to inhibit Ebola infection

The Ebola virus experiments were conducted in the Biosafety level 4 (BSL4) facility at the United States Army Medical Research Institute of Infectious Diseases (USAMRIID). Dendritic cells from four human donors were infected with Ebola Zaire species, strain Mayinga and then genome-wide gene expression analysis with the Affymetrix Human Genome U133 Plus 2.0 Array (Santa Clara, CA, USA) was performed prior to infection and at 30, 60, and 120 min after infection. This gene expression microarray data is available for download from the help section of the L1000CDS^2 tool. The differentially expressed genes at each time point were computed comparing each time point with the control. The three signatures of the differentially expressed genes were then queried with the L1000CDS^2 tool to prioritize drugs and small molecules that can potentially mimic or reverse the gene expression state of the Ebola-infected cells. Top ranked drugs were selected for experimental validation. To test each drug HeLa cells were pretreated with each top ranked small molecule and then infected with Ebola at a MOI of five for 48 h. Ebola-infected cells were then stained for viral antigen and analyzed on an Opera, confocal high-content imaging platform. Image analysis was performed using the acapella software. Using this system we were able to account for cell viability by nuclear (Hoechst) and cytoplasmic (CellMask) stains (Waltham, MA, USA). To determine whether kenpaullone and other small molecules can inhibit Ebola infection in a dose-dependent manner, HeLa cells were pretreated with each drug in a dose range and then infected with Ebola at an MOI of five for 48 h. The same experimental setup was applied to evaluate human foreskin fibroblast cells.

ACKNOWLEDGEMENTS

This work is supported by NIH grants: R01GM098316, U54HG008230 and U54CA189201 to A.M. The Joint Science and Technology Office for Chemical and Biological Defense (JSTO-CBD) of the Defense Threat Reduction Agency (DTRA) to S.B. A.D.R. is supported by T32HL007824.

COMPETING INTERESTS

The authors declare no conflict of interest.

REFERENCES

1. Stegmaier, K. *et al.* Gene expression–based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nat. Genet.* **36**, 257–263 (2004).
2. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
3. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
4. Clark, N. R. *et al.* The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics* **15**, 79 (2014).
5. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420 (Springer, 2005).
6. Anders, S. Analysing RNA-Seq data with the DESeq package. *Mol. Biol.* **43**, 1–17 (2010).
7. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
8. McLauchlan, H., Elliott, M. & Cohen, P. The specificities of protein kinase inhibitors: an update. *Biochem. J.* **371**, 199–204 (2003).
9. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
10. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
11. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
12. Consortium E. P. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**, 636–640 (2004).
13. Lachmann, A. & Ma'ayan, A. KEA: kinase enrichment analysis. *Bioinformatics* **25**, 684–686 (2009).
14. Chen, E. Y. *et al.* mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics* **28**, 105–111 (2012).
15. Isik, Z., Baldow, C., Cannistraci, C. V. & Schroeder, M. Drug target prioritization by perturbed gene expression and network information. *Sci. Rep.* **5**, 17417 (2015).
16. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

17. Yen, J. Y. *et al.* Therapeutics of ebola hemorrhagic fever: whole-genome transcriptional analysis of successful disease mitigation. *J. Infect. Dis.* **204**, S1043–S1052 (2011).
18. Wahl-Jensen, V. *et al.* Ebola virion attachment and entry into human macrophages profoundly effects early cellular gene expression. *PLoS Negl. Trop. Dis.* **5**, e1359 (2011).
19. Rasmussen, A. L. *et al.* Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance. *Science* **346**, 987–991 (2014).
20. Panchal, R. G. *et al.* Reduced levels of protein tyrosine phosphatase CD45 protect mice from the lethal effects of Ebola virus infection. *Cell Host Microbe* **6**, 162–173 (2009).
21. Rubins, K. *et al.* The temporal program of peripheral blood gene expression in the response of nonhuman primates to Ebola hemorrhagic fever. *Genome Biol.* **8**, R174 (2007).
22. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
23. Gu, H., Ding, L., Xiong S.-d., Gao X.-m. & Zheng, B. Inhibition of CDK2 promotes inducible regulatory T-cell differentiation through TGFβ-Smad3 signaling pathway. *Cell Immunol.* **290**, 138–144 (2014).
24. Zhong, Z., Wen, Z. & Darnell, J. Stat3: a STAT family member activated by tyrosine phosphorylation in response to epidermal growth factor and interleukin-6. *Science* **264**, 95–98 (1994).
25. Dugo, L. *et al.* in *Sepsis: New Insights, New Therapies: Novartis Foundation Symposium 280* 128–146 (Wiley Online Library, 2007).
26. Bertsch, S., Lang, C. H. & Vary, T. C. Inhibition of GSK-3β activity with lithium in vitro attenuates sepsis-induced changes in muscle protein turnover. *Shock (Augusta, GA)* **35**, 266 (2011).
27. Claus, C. & Liebert, U. G. A renewed focus on the interplay between viruses and mitochondrial metabolism. *Arch. Virol.* **159**, 1267–1277 (2014).
28. Bowen, E. *et al.* The effect of interferon on experimental Ebola virus infection in rhesus monkeys. *Interferon* **5**, 5 (1978).
29. Kunick, C., Lauenroth, K., Leost, M., Meijer, L. & Lemcke, T. 1-Azakenpaullone is a selective inhibitor of glycogen synthase kinase-3β. *Bioorg. Med. Chem. Lett.* **14**, 413–416 (2004).
30. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **573**, 83–92 (2004).
31. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091–D1097 (2014).
32. Wang, Z., Clark, N. R. & Ma'ayan, A. Dynamics of the discovery process of protein-protein interactions from low content studies. *BMC Syst. Biol.* **9**, 26 (2015).
33. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091–D1097 (2014).
34. Schomburg, K., Ehrlich, H. C., Stierand, K. & Rarey, M. From structure diagrams to visual chemical patterns. *J. Chem. Inf. Model.* **50**, 1529–1535 (2010).

Supplemental Information accompanies the paper on the *Systems Biology and Applications* website (http://www.nature.com/npjsba)