Supplementary Materials

Breaking the paradigm: Dr. Insight empowers signature-free, enhanced drug repurposing

Jinyan Chan[1,2], Xuan Wang[1], Jacob Turner[3], Nicole Baldwin[1], Jinghua Gu[1,*]

[1] Baylor Scott & White Research Institute, 3310 Live Oak St, Dallas, TX 75204, USA., [2] Institute of Biomedical Studies, Baylor University, One Bear Place #97224, Waco, TX 76706, USA., [3] Department of Mathematics and Statistics, Stephen F. Austin State University, PO Box 13040 SFA Station, Nacog- doches, TX 75962, USA.

*To whom correspondence should be addressed.

Table of Contents

Supplementary Materials

S1. Data resources

S1.1 Data sets used in this study

Reference drug profiles of the CMap drug perturbation dataset (build 02) were downloaded from the CMap website (https://portals.broadinstitute.org/cmap/). This dataset consists of 6100 ranked gene lists (i.e. drug instances) that are derived from the differential gene expression analysis between drug-treated and untreated human cell lines. These drug instances represent 3587 drug treatment sets (see S3.2 for the definition of drug treatment set) [1] that cover 1309 distinct drugs.

Three gene expression datasets from breast and prostate cancer samples, and two gene expression datasets from non-cancer diseases: systemic lupus erythematosus (SLE) and hepatitis B infection (HBV), were used in this study. We downloaded the breast cancer and prostate cancer level-3 RNA-Seq data from The Cancer Genome Atlas (TCGA) Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). The microarray-based prostate cancer (GSE3325) gene expression dataset [2], SLE gene expression dataset (GSE65391) and HBV dataset (GSE69590) were obtained from the Gene Expression Omnibus (GEO) database. Following Sinha *et al*.[3], and Borcherding *et al*. [4] a two-sample t-test was performed on log-transformed TCGA level-3 normalized count data to analyze the differential expression between disease and normal samples. For the microarray data from GEO, we also performed two-sample t-test between disease and non-disease samples. For the SLE dataset, we only included the samples from batch 2 (there are two different batches in this dataset) for the analysis. Details of the three cancer datasets are listed in Table S3. The two TCGA RNA-seq datasets contain the expression values of each gene (gene symbol), while the GEO prostate cancer dataset and HBV dataset were assayed by Affymetrix Human Genome U133 Plus 2.0 Array and SLE dataset by Illumina HumanHT-12 v3 Array. When querying methods such as CMap and sscMap, the gene symbols or probe IDs in all datasets were converted to Affymetrix probe IDs because these software require the input gene signatures to be probe IDs. On the other hand, Cogena and Dr. Insight use gene symbols as input, which required us to convert different probe IDs in the GEO datasets to gene symbols before analysis.

S1.2 Ground-truth disease-reversing drugs

For breast cancer and prostate cancer, we used FDA-approved drugs and drugs that were in advanced clinical trials [5] as ground-truth drugs for performance validation. The breast cancer ground-truth drug set includes 195 drug treatment sets that represent 72 distinct drugs. The prostate cancer drug set contains 155 drug treatment sets that cover 55 distinct drugs.

For the two additional non-cancer datasets, we collected drugs that were in advanced clinical trials from https://clinicaltrials.gov. To be more specific, we searched the disease names (systemic lupus erythematosus and hepatitis B infection individually) in https://clinicaltrials.gov, and selected drugs that were in phase 3 and 4 clinical trials with either "completed" or "active" status. For SLE data, we also included the drugs that were well established to treat SLE from https://resources.lupus.org/entry/medications-used-to-treat-lupus. Finally, we merged these drugs with the total drug list from the CMap database as our ground truth drug set. Lists of the ground-truth drugs are listed in Table S4.

S1.3 Ground-truth drug target genes

We collected the documented drug target genes from STITCH and CTD databases to evaluate if the identified CEGs were over-represented by reported drug targets. The drug-target interactions acquired from the STITCH database contain only undirected interactions between drugs and target genes, whereas the information from CTD database is directed (i.e., each interaction in the database is annotated as up-regulation or down-regulation).

S2. Proofs of theorems

Here we prove that: Theorem 1, given two independent rank variables X and Y with the same range $\{1,2,\dots,N\}$, the statistic $R = \min\left(\frac{X}{N}, \frac{Y}{N}\right)$ has the cumulative distribution function (CDF) of $I_r(a=1, b=2)$, where $I_r(a,b)$ is the regularized incomplete Beta function; and Theorem 2: given two rank variables X and Y with the same range $\{1,2,\dots,N\}$, $R = \max\left(\frac{X}{N}, \frac{Y}{N}\right)$ has the cumulative distribution function (CDF) of $I_r(a=2, b=1)$, where $I_r(a,b)$ is the regularized incomplete Beta function.

Theorem 1 Proof: We first calculate the CDF for R as follows:

$$F_R(r) = P(R \le r) = 1 - P\left(\frac{X}{N} > r \cap \frac{Y}{N} > r\right) = 1 - \left(\frac{N - Nr}{N}\right)^2 = 2r - r^2, \qquad r \in \{\frac{1}{N}, \frac{2}{N}, \dots, \frac{N}{N}\}$$

Also, the regularized incomplete Beta function is defined as:

$$I_r(a, b) = \frac{\mathrm{B}(r; a, b)}{\mathrm{B}(a, b)},$$

where $\mathrm{B}(r; a, b) = \int_0^r t^{a-1}(1-t)^{b-1}\, dt$ is the incomplete Beta function. $\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the complete Beta function, where $\Gamma(n) = (n-1)!$ is the Gamma function for positive integer n. If we substitute $a = 1, b = 2$ into $I_r(a, b)$ we have:

$$I_r(1,2) = \frac{\mathrm{B}(r; a, b)}{\mathrm{B}(a, b)} = \frac{\int_0^r t^0(1-t)^1\, dt}{\frac{0!\,1!}{2!}} = 2r - r^2 = F_R(r).$$

Theorem 2 Proof: We first calculate the CDF for R as follows:

$$F_R(r) = P(R \le r) = P\left(\frac{X}{N} \le r \cap \frac{Y}{N} \le r\right) = \left(\frac{Nr}{N}\right)^2 = r^2, \qquad r \in \{\frac{1}{N}, \frac{2}{N}, \dots, \frac{N}{N}\}$$

Similarly, the regularized incomplete Beta function is given by:

$$I_r(2,1) = \frac{\mathrm{B}(r; 2,1)}{\mathrm{B}(2,1)} = \frac{\int_0^r t^1(1-t)^0\, dt}{\frac{1!\,0!}{2!}} = r^2 = F_R(r).$$

S3. Selection of CEG cutoff parameter

To evaluate how different p-value thresholds of CEGs affect the performance of Dr. Insight, we tested a spectrum of p-values parameters from 0.0001 to 0.2 to select CEGs. Same as the performance evaluation strategy used in the main article, we performed enrichment analysis to assess whether the identified drug candidates by CEG threshold had statistically significant overlap with the ground-truth drugs. We applied this evaluation procedure to all five datasets

(three cancer datasets that were originally reported in the manuscript and two additional non-cancer datasets).

The bar plots in Figure S1 give the log-transformed p-values of the enrichment test results from six different CEG cutoffs (0.0001, 0.001, 0.01, 0.05, 0.1, 0.2; red bars). In addition, we also added the enrichment scores of the CMap method (blue bars) side-by-side with Dr. Insight's results as references. First of all, as shown by the bar plots, Dr. Insight apparently outperformed CMap in most scenarios regardless of the CEG cutoffs applied. For Dr. Insight itself, there is a clear trend that the performance of Dr. Insight consistently degraded when the cutoff was set to be either "very significant" (0.0001) or "insignificant" (0.2), while the peak performance of the repurposing algorithm was always achieved when the cutoff was set close to 0.05. This observation is well expected, where the behavior of Dr. Insight should highly correlate with the amount of useful information incorporated into the model. In other words, when setting the cutoff to a very stringent level such as 0.0001, we only utilize a small portion of top CEGs with highest differential expression, yet we exclude many significant CEGs with decent p-values but are not ranked on the top. Taking the TCGA breast cancer data as an example, when applying a cutoff of 0.001, we only had 3 CEGs, compared to 564 CEGs returned by the cutoff of 0.05 for drug fulvestrant. As a result, Dr. Insight detected fulvestrant as a significant drug with a drug p-value of 2.57e-4 for the 0.05 cutoff, yet it remained insignificant for the 0.001 cutoff (drug p-value=0.23). On the other hand, when a looser threshold is used (e.g., 0.2), it introduces more noise into the model by including non-significant CEGs, hence negatively impacts the performance of the method.

Our suggested cutoff (p-value = 0.05) had consistent leading performance (at least top 2) across all five datasets. In some cases, we observe that a p-value cutoff of 0.01 and 0.1 may result in slightly better enrichment, yet their difference compared to the 0.05 cutoff was very minimum. For instance, cutoff=0.01 yielded an enrichment p-value of 1.09e-4 in the TCGA prostate cancer dataset, compared to the enrichment p-value of 1.25e-4 by the suggested 0.05 cutoff; in the HBV study, cutoff=0.1 resulted in an enrichment p-value of 1.22e-3, compared to 1.56e-3 given by the 0.05 cutoff. All taken together, the above empirical evidences suggest that the recommended 0.05 cutoff serves as a very robust threshold for selecting CEGs in Dr. Insight. We used this parameter for CEG selection throughout the experiments in the manuscript. Meanwhile, in the Dr. Insight R package, we provided the parameter "CEG.threshold" in the function "drug.ident" so that users adjust the CEG threshold when needed.

S4. Gene signatures and drug treatment sets

S4.1 Gene signatures

To evaluate the drug repurposing performance of CMap, sscMap and NFFinder, which do not have specific recommendations for the size of query gene signatures, we used gene lists of varying sizes (50, 100, 200, 300, 400, 600, 800 and 1000 Affymetrix probes) for query. These gene signatures were composed of top- and bottom-ranked most differentially expressed genes. The size range was determined based on numerous existing CMap applications, where small (40, 100 or 150 probes [6-8]), medium (230 or 406 probes [9, 10]), and larger sizes (870 to 1000 probes [10, 11]) were used. We set 1000 probes as the upper limit for analysis in this study because it is the biggest signature size allowed for the CMap software. The input for Cogena was the original gene expression data of the signature genes (rather than only the gene lists). Because the Cogena paper has the recommended criteria of FDR ≤ 0.05 and log-fold-change ≥ 1, so we selected the gene signature with these criteria to evaluate the performance of Cogena in both simulation and real data studies.

S4.2. Drugs are studied at treatment set level

We define a drug treatment set as a set of drug instances collected from one cell line that is treated by one drug. For example, “tamoxifen_MCF7” drug treatment set contains four drug instances that were obtained from tamoxifen perturbed MCF7 cell line. We take drug treatment set as the unit for our drug study. The drug treatment set is referred to as “drug” for simplicity in the results description.

S5. Realistic simulation of disease-drug associations

The procedure of the realistic simulation is illustrated in Figure S3, and it includes the following main steps:

(1) Select synthetic positive drug set based on clustering analysis.

We first generated the prototype list (PRL) of each individual drug by merging all of its instances in CMap into one ranked gene list [12]. Clustering analysis was performed on PRLs of all the drugs and a cluster of highly correlated drugs were selected as the synthetic positive drug set G for simulation analysis.

To be more specific, we first merged all 6100 drug instances into the PRLs of 3587 drug treatment sets (see definition of “drug treatment set” in S3.2) with the aggregation algorithm

introduced by Iorio et al. [12]. All the drug treatment sets were then clustered by their pairwise Spearman's correlation values. A cluster C containing 697 drug treatments with high within-cluster similarity was chosen as the drug candidate pool for generating synthetic positive drug set G (Figure S3-1). For each realization of the simulation, we randomly selected 20 drug treatments from C as the ground-truth drug set G. The remaining drugs in CMap were referred to as negative drug set $\bar{G}$. The rank lists of synthetic positive drugs in G were then merged to generate a prototype PRL list designated as P_G.

(2) Learn ground-truth disease-drug association patterns from FDA-approved drugs.

To realistically mimic disease-drug associations for synthetic ground-truth drugs, we first investigated the rank correlations between real cancer datasets and the gene expression of their corresponding FDA-approved drugs. A sub-set of the FDA drugs with higher correlations with the disease data were selected as reference positive drug set g. The rank list of disease data was denoted as L_{ref}. To train the rank correlation pattern $R_{disease\leftrightarrow drug}$ from FDA drugs, we used a Mote Carlo procedure to randomly sample a new rank vector P_g from the pooled rank lists of drug set g. A non-parametric model of disease-drug association was given by: $L_{\mathrm{ref}} = R_{disease\leftrightarrow drug}(P_g)$.

(3) Generate synthetic disease rank list L_G.

Based on the inferred disease-drug association pattern $R_{disease\leftrightarrow drug}$ from step (2), we calculated the synthetic disease rank list $L_G = R_{disease\leftrightarrow drug}(P_G)$. L_G was used as the input for the simulation study and drug set G was the ground-truth drug set.

Repeated experiments were performed (n=10) to account for the variability in the simulation study and the average performance was reported.

S6. Additional information for performance comparison for drug detection

We evaluated drug predictability of each method by accessing its ability to correctly predict ground-truth drug treatment sets. The drug treatment set predicting p-values were directly reported by CMap, sscMap and Dr. Insight. However, for NFFinder and Cogena, only instance-level p-values were reported. Hence, we use the minimum p-value of all the instances within a particular drug treatment set as its overall statistical significance.

In simulation studies, receiver operating characteristic (ROC) curve was used to compare the drug prediction performance from different repurposing methods. To focus on negative

connectivity (drug treatment effect), p-values associated with positive connectivity by CMap and sscMap were set to 1. For CMap, sscMap and NFFinder, multiple gene signatures with different sizes were used for the query, which yielded multiple ROC curves. For each of these signature-based methods, the average of eight ROC curves that corresponded to eight query signatures sizes was plotted to show the overall performance. The range of performance for different signatures was shown by the shaded area in Figure S4.

When studying real cancer datasets, we compared the drug prediction performance of these methods by evaluating whether the proposed drug treatment candidates ($p \leq 0.05$) by each method had statistically significant overlap with ground-truth drugs (Fisher's exact test). The number of detected significant drugs from each method and the number of successfully predicted ground-truth drugs are listed in Table S5.1 – Table S5.11.

S7. Additional information for drug target prediction

Another major goal of the study is to validate that the concordantly expressed genes (CEGs) identified by Dr. Insight are better surrogates for drug target prediction compared to differentially expressed genes (DEGs). To this end, we evaluated the enrichment of ground-truth drug targets collected from public drug interaction databases, such as STITCH and CTD, for both CEG and DEG-based methods. As is demonstrated in the main article, for each of the identified drugs by Dr. Insight, we used Fisher's exact test to assess if the CEGs of this drug were over-represented by known target genes collected in the STITCH or CTD databases. If a statistical significant p-value (e.g., 0.05) is obtained, the corresponding drug is deemed "enriched" in known drug-target interactions ("enriched drugs" for short). Similarly for the DEG-based methods, i.e., CMap, sscMap, and NFFinder, enrichment analysis of known drug targets was performed on the differentially expressed genes identified from the disease data. Finally, the percentage of enriched drugs out of total identified drugs, for both CEG and DEG-based methods, were calculated and compared in Figure 5, Figure S6 and Figure S7.

It is worth noting that drug targets from the CTD database have directions (i.e., whether a gene is up-regulated or down-regulated after drug perturbation). Therefore, more specific evaluation of target prediction can be achieved using drug targets from the CTD database by taken into consideration the direction information. In other words, we evaluated if the up-regulated CEGs/DEGs were particularly enriched with drug up-regulated targets, and vice versa for the down-regulation.

The numbers and the percentages of enriched drugs for Dr. Insight are listed in Table S6 and Table S7, including results from the three cancer datasets. The percentages of the CMap-proposed drugs whose DEGs are enriched with true targets were shown in the main text (Figure 5). Figure S6 and Figure S7 show the bar plots of the percentages of the sscMap and NFFinder-proposed drugs whose DEGs are enriched with true targets. In both figures, the percentage of the Dr. Insight-proposed drugs whose CEGs are enriched with true targets were also included for comparison.

S8. Systematic comparison of drug actionable targets in pathways

The pathway analysis results of Dr. Insight are tightly connected to the selected CEGs. For a given drug, a significant pathway should have a greater outlier-sum (a function of CEGs and their corresponding z-scores, Eq. (5) of the main text) than (1) the outlier-sum of the same pathway for other drugs and (2) the outlier-sum of the rest of the pathways for the same drug. We have shown in the main manuscript (3.4 CEGs significantly improve drug target prediction and Figure 4) that CEGs are better proxy for drug target prediction, compared to DEGs used by other signature-based methods [13-16]. The fact that our pathway analysis highly depends on the selected CEGs will hence inherit their predictive power. In other words, it is rational to expect that the identified pathways by Dr. Insight will be more enriched in "drug actionable targets" (i.e., genes with reversed expression between disease and drug-perturbed data) compared to DEG-based methods.

To verify the above point, we systematically evaluated the percentage of "drug actionable targets" in pathway analysis of TCGA breast cancer data using both CEG and DEG-based methods. We define a drug actionable target as a gene with significantly reversed expression (ranked top/bottom 10%) in both disease and drug-perturbed data. Multiple DEG signatures with varying sizes were used for CMap. For each drug identified, we performed a hypergeometric test on all DEG signatures to identify significantly enriched PID pathways. Similar as we did for the CEGs, we took the DEGs within these significant pathways and measured the percentage of drug actionable targets. From Table S8, we see that CEGs (average 740 genes; minimum 234 genes; maximum 1099 genes) gives an average of more than 40% enrichment in genes that are consistently ranked in the top/bottom of both disease and drug data, compared to less than 10% of enrichment by DEG-based method. These results show that the aberrant regulation of the pathways can be effectively normalized by drugs proposed by Dr. Insight. The DEG-based

method, on the other hand, “greedily” searches for pathways that have the highest expression change in the disease data, yet their reversibility to drug treatment is compromised.

S9. Robustness evaluation for drug detection

To address the robustness of Dr. Insight against noise in the data, we performed a sensitivity analysis by adding different levels of noise to the disease dataset and evaluated the performance of three methods: Dr. Insight, simple inverse correlation and the original CMap method. We selected one cancer dataset: prostate cancer data from TCGA, and one non-cancer dataset: systemic lupus erythematosus dataset as examples, to demonstrate how much noise can be tolerated by each method.

To simulate datasets with controlled noise effects, we first used the inverse cumulative distribution function (ϕ^{-1}) for normal distribution to transform the original scaled gene ranks r/N into continuous numbers x as $x = \phi^{-1}\left(\frac{r}{N}\right)$, where N is the total number of genes and r is the rank of each gene. We then generated the noise ε from a standard normal distribution. We combined different percentages of the noise data and disease data to generate “noise-corrupted” input:

$$y = \eta x + (1-\eta)\epsilon,$$

where $\eta = 0\%, 20\%, 60\%, 80\%, 90\%,$ denoting a spectrum of noise settings. Finally, we sort y and generate the new rank r' for downstream analysis. To assess how well each method tolerates noise, we used enrichment test to measure whether the drugs repurposed by each method at a given noise level have significant overlap with the ground truth drugs for the disease. Each noise setting was simulated for 20 times and the p-values of the enrichment tests were summarized. Figure S10 gives the median of log-transformed enrichment p-value of the three methods at different noise levels.

Figure S10 shows that in both prostate cancer and SLE data, Dr. Insight consistently achieved the highest enrichment p-values among all three methods, validating its robust and superior performance to noisy datasets. In the prostate cancer simulation study, both Dr. Insight and the inverse correlation method were robust against added noise, where they produced significant enrichment (p-value <0.05, indicated by black horizontal line) even with 90% of noise. As a contrast, the original CMap method was much more sensitive to noise, most likely due to that CMap only utilizes a limited amount of data (i.e., a set of selected signature genes from the disease data) for drug repurposing. Similar results have been observed on the SLE dataset,

where Dr. Insight and inverse correlation showed robustness to as much as 60% noise, whereas CMap can hardly produce any significant results even without added noise.

S10. Software and reproducibility of breast cancer study

The R package of Dr. Insight can be freely downloaded from https://cran.r-project.org/web/packages/DrInsight/index.html. Our package depends on two existing packages “igraph” [13] and “qusage” [14]. In our vignette, we have given detailed instructions to reproduce our TCGA case study (https://cran.r-project.org/web/packages/DrInsight/vignettes/my-vignette.html).

Supplementary Tables

Table S1. Signature selection for CMap analysis from the review paper by Musa *et al*. [15]

Disease	Dataset	Signature criteria	Signature size
prostate cancer	Celastrol- and gedunin-treated cell lines (GSE5505 and GSE5508)	Criteria unknown; signatures validated by GE-HTS bead-based assay	27
T-ALL	Human and mouse T-ALL cell lines (GSE12948, GSE8416 and GSE14618)	p-value ≤ 0.05	150
Gastric cancer	Yonsei gastric cancer (GSE13861)	p-value ≤ 0.001 & fold change ≥ 2	1000
CNS injuries	Human MCF7 breast adenocarcinoma (GSE34331)	p-value ≤ 0.05 & fold change ≥ 1.5	21
GBM	GSE4290	p-value ≤ 0.0001 & fold change ≥ 4	406
	GSE7696		270
	GSE14805		870
	GSE15824		111
	GSE16011		1000
Stem cell leukemia (SCL)	hESCs cell lines (GSE54508)	p-value ≤ 0.01 & fold change ≥ 2, then top and bottom100 genes	200
Myelomatosis	Human myeloma cell lines (GSE14011)	FDR ≤ 0.25 & fold change ≥ 1.5	38
AML	AML data (GSE7538)	Top and bottom 75 probes	150
MLL-rearranged infant ALL	GSE32962	FDR ≤ 0.05	50
Ovarian cancer	GSE82007	p-value ≤ 0.001	60

Table S2. Comparison of connectivity-mapping based drug repurposing methods. The "Method name" column gives a list of representative methods that use CMap drug profiles or other drug perturbed expression data. The methods listed in the brackets are variations of the original method with technical improvements. For example, methodology-wise, gCMAP is very similar to CMap, except that it implements some existing gene set enrichment methods to evaluate the connectivity and aggregates the mapping tool into an R package; QUARrATiC borrows the sscMap methodology but extends the reference drug perturbation datasets from CMap data to The Library of Integrated Network-Based Cellular Signatures (LINCS) data. Dr. Insight is the only method that does not belong to the category of "two-step, signature-based" repurposing framework, while at the same time, it is equipped with the capability to predict drug targets and perform functional analysis at pathway/network level.

Method name	Drug repurposing		Drug mechanism study	
	Signature-based two-step model?	Fixed signature size?	Target prediction?	Pathway and functional analysis?
CMap (gCMAP [16], DMAP [17], DvD [18])	Y	Y	N	N
sscMap [19] (QUADrATiC [20])	Y	Y	N	N
CDA [21]	Y	Y	N	Y
Shigemizu, et al. 2012 [22] (Chen, et al. 2016 [5])	Y	N	N	Y
NFFinder [23]	Y	Y	N	N
Cogena [24]	Y	Y	N	Y
Wen, et al. 2016 [25]	Y	N	N	N
Dr. Insight	N	N	Y	Y

Table S3. Disease datasets.

Datasets	TCGA breast cancer (BRCA)	TCGA prostate cancer (PRAD)	GEO prostate cancer (PRAD)	Systemic lupus erythematosus (SLE)	Hepatitis B infection (HBV)
Data type	RNAseq data	RNAseq data	Microarray data	Microarray data	Microarray data
# case samples	1099	498	7	118	3
# control samples	111	52	6	40	3
# genes used in analysis	11067	11067	12994	12994	12994

Table S4. **List of ground-truth drugs that are used as benchmarks in our study.** For breast cancer and prostate cancer, FDA-approved drugs have been highlighted in red. The rest of the drugs are in advanced clinical trials [5].

Breast cancer	Prostate cancer	Systemic lupus erythematosus	Hepatitis B infection
tamoxifen	estropipate	aspirin	estradiol
exemestane	flutamide	acetaminophen	tacrolimus
raloxifene	megestrol	ibuprofen	zidovudine
fulvestrant	aminoglutethimide	naproxen	levonorgestrel
paclitaxel	nilutamide	prednisolone	ribavirin
methotrexate	diethylstilbestrol	methylprednisolone	
letrozole	cyproterone	indomethacin	
doxorubicin	mitoxantrone	nabumetone	
vinblastine	rofecoxib	celecoxib	
camptothecin	genistein	cyclophosphamide	
vorinostat	methylprednisolone	methotrexate	
simvastatin	lovastatin	azathioprine	
etoposide	tacrolimus	hydroxychloroquine	
irinotecan	celecoxib	chloroquine	
estradiol	fluvastatin	heparin	
lovastatin	camptothecin	warfarin	
prednisolone	doxorubicin	belimumab	
sirolimus	papaverine	ACTH	
metformin	vorinostat	adrenocorticotropic	
naproxen	simvastatin		
dexamethasone	sirolimus		
trifluridine	finasteride		
lidocaine	prednisolone		
aminoglutethimide	testosterone		
alvespimycin	metformin		
ciprofloxacin	azacitidine		

azacitidine naltrexone tacrolimus gefitinib ribavirin estriol chlorhexidine testosterone omeprazole loperamide pentoxifylline bisoprolol doxycycline digoxin fluvoxamine captopril ranitidine thalidomide tanespimycin minocycline gabapentin celecoxib methylprednisolone prednisone ramipril hydralazine clonidine bupivacaine diclofenac imatinib mifepristone propranolol riluzole ondansetron melatonin nimesulide chloroquine sulindac propofol prochlorperazine dacarbazine lisinopril metoprolol decitabine	gefitinib thalidomide hydrocortisone gabapentin dexamethasone fulvestrant niclosamide leflunomide tamoxifen dextromethorphan amantadine paclitaxel exemestane theophylline mifepristone ciprofloxacin pioglitazone tanespimycin ranitidine quercetin etoposide isotretinoin omeprazole propofol prednisone diphenhydramine colchicine bupivacaine		

ifosfamide cyclobenzaprine			

Table S5.1. **CMap drug identification results with TCGA BRCA data**

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	56	52	55	67	74	74	71	56
# identified ground-truth drug treatments	5	6	3	6	5	6	6	5
Enrichment p value	0.214	0.074	0.617	0.182	0.417	0.246	0.218	0.218

Table S5.2. **CMap drug identification results with TCGA PRAD data**

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	52	83	65	66	80	73	76	71
# identified ground-truth drug treatments	8	11	7	7	9	10	8	9
Enrichment p value	0.002	0.001	0.021	0.022	0.007	0.001	0.016	0.003

Table S5.3. CMap drug identification results with GEO PRAD data

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	62	57	61	58	53	48	66	65
# identified ground-truth drug treatments	6	6	7	6	4	4	7	3
Enrichment p value	0.049	0.035	0.015	0.037	0.194	0.151	0.022	0.539

Table S5.4. **sscMap drug identification results with TCGA BRCA data**

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	6	19	659	998	1185	1316	1404	1436
# identified ground-truth drug treatments	0	1	31	45	61	63	72	77
Enrichment p value	1.000	0.674	0.911	0.980	0.866	0.972	0.902	0.793

Table S5.5. sscMap drug identification results with TCGA PRAD data

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	8	18	100	177	202	381	653	810
# identified ground-truth drug treatments	1	4	9	11	14	17	21	25
Enrichment p value	0.298	0.006	0.027	0.141	0.052	0.483	0.954	0.983

Table S5.6. sscMap drug identification results with GEO PRAD data

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	4	8	17	15	10	11	26	29
# identified ground-truth drug treatments	1	2	4	3	1	0	1	0
Enrichment p value	0.162	0.044	0.005	0.025	0.357	1.000	0.684	1.000

Table S5.7. NFFinder drug identification results with TCGA BRCA data

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	329	285	623	782	651	854	1007	1069
# identified ground-truth drug treatments	26	24	40	45	43	54	66	62
Enrichment p value	0.052	0.032	0.227	0.508	0.161	0.213	0.103	0.471

Table S5.8. NFFinder drug identification results with TCGA PRAD data

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	347	592	548	717	529	719	842	783
# identified ground-truth drug treatments	18	28	32	34	32	38	42	38
Enrichment p value	0.238	0.329	0.041	0.298	0.026	0.096	0.161	0.231

Table S5.9. NFFinder drug identification results with GEO PRAD data

Gene signature sizes	50	100	200	300	400	600	800	1000
# identified drug treatments	191	159	249	207	216	323	292	372
# identified ground-truth drug treatments	16	17	20	18	20	27	20	28
Enrichment p value	0.007	3.98E-4	0.005	0.003	0.001	0.001	0.025	0.002

Table S5.10. Cogena drug identification results

Cogena	TCGA BRCA	TCGA PRAD	GEO PRAD
# identified drug treatments	335	98	27
# identified ground-truth drug treatments	30	5	2
Enrichment p value	0.008	0.418	0.327

Table S5.11. Dr. Insight drug identification results

Dr. Insight	TCGA BRCA	TCGA PRAD	GEO PRAD
# identified drug treatments	70	69	53
# identified ground-truth drug treatments	15	11	11
Enrichment p value	5.96E-6	1.25E-4	1.32E-5

Table S6. Drugs identified by Dr. Insight whose CEGs are enriched with STITCH drug targets

	# enriched drugs	% of enriched drugs
TCGA BRCA dataset	15	44%
TCGA PRAD dataset	10	29%
GEO PRAD dataset	11	42%

Table S7. Drugs identified by Dr. Insight whose CEGs are enriched with CTD drug targets

	# up targets enriched drugs	% of up targets enriched drugs	# down targets enriched drugs	% of down targets enriched drugs
TCGA BRCA dataset	15	47%	12	38%
TCGA PRAD dataset	16	47%	11	31%
GEO PRAD dataset	19	63%	14	50%

Table S8. Percentage of pathway CEGs/DEGs that are consistently ranked at the top/bottom (inversely expressed) 10% of drug and disease data.

Gene type	Number of drugs identified	Number of pathways identified	Percentage of consistently top/bottom ranked genes
CEG	10	31	43.0%
DEG (50 probes)	2	2	0.96%
DEG (100 probes)	1	1	2.60%
DEG (200 probes)	0	5	3.73%
DEG (300 probes)	7	5	5.01%
DEG (400 probes)	6	8	5.24%
DEG (600 probes)	4	8	5.19%
DEG (800 probes)	3	8	5.15%
DEG (1000 probes)	6	9	5.70%

Supplementary Figures

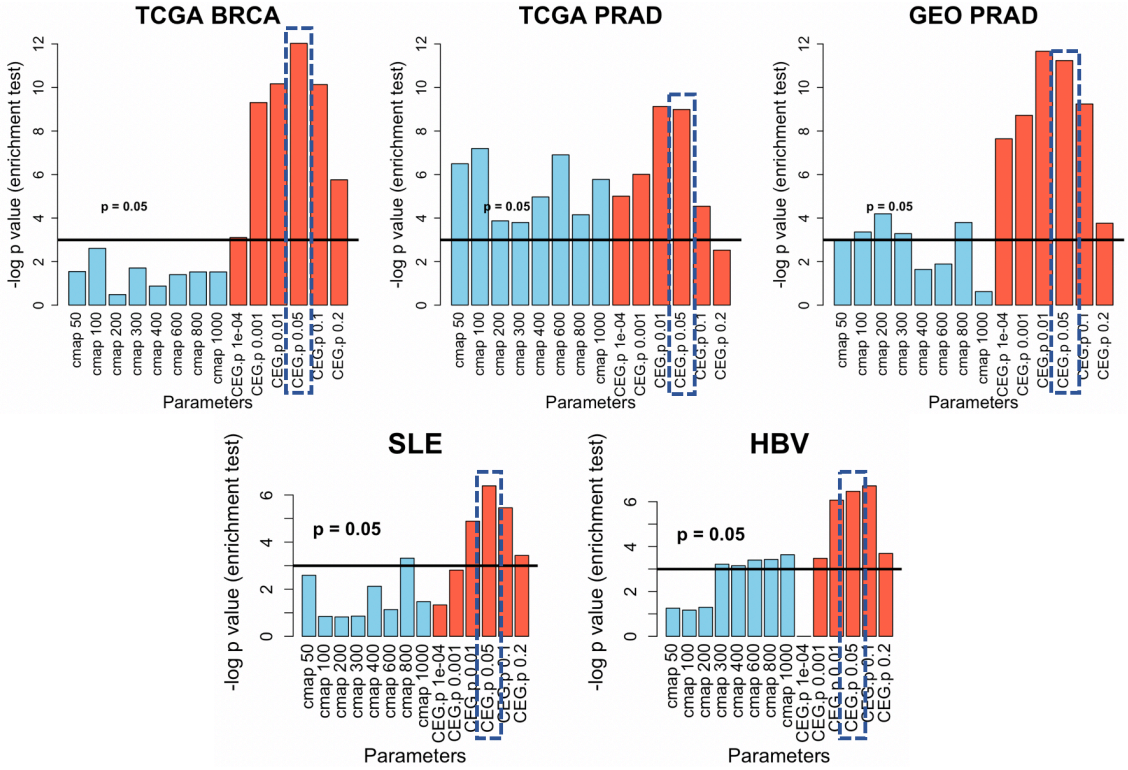

Figure S1. Dr. Insight performance with different CEG cutoff. The x axis shows the different CMap signature size parameters, and CEG p-value cutoff parameters: 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2. The y axis is the -log p-value of the enrichment test results of candidate drugs. The CEG parameter used in the manuscript is in the dotted boxes. TCGA BRCA: breast cancer dataset from TCGA. TCGA PRAD: prostate cancer dataset from TCGA. GEO PRAD: prostate cancer dataset from GEO. SLE: systemic lupus erythematosus dataset. HBV: Hepatitis B virus infection dataset.

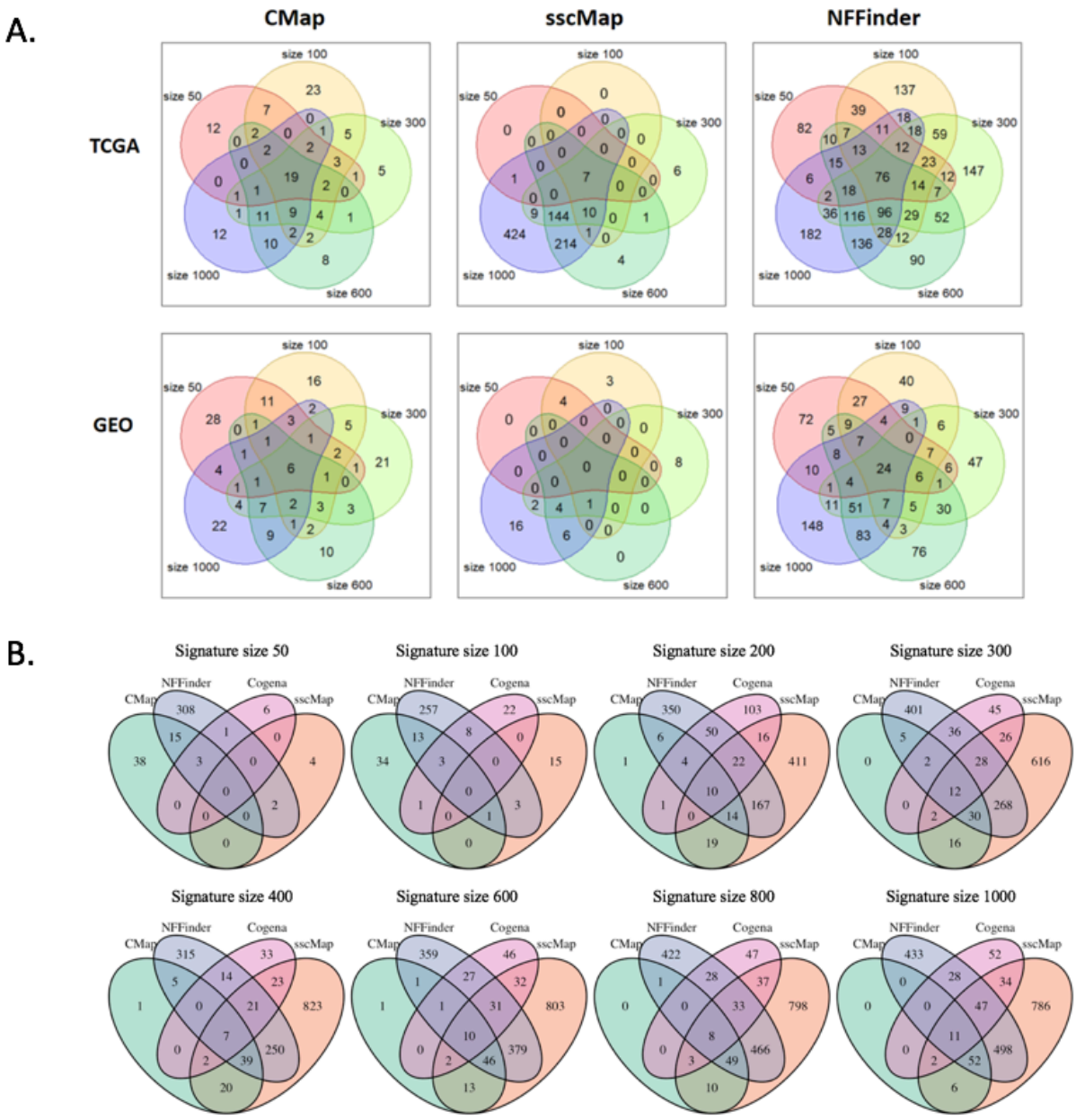

Figure S2. Drug identification results comparison of different methods with various signature sizes. A. Drug identification results comparison among five different signature sizes on two prostate cancer datasets (upper panel: TCGA data; lower panel: GEO data). The Venn Diagrams show the agreement of drug identification results among the five signatures sizes (50, 100, 300, 600, and 1000 probes). Each Venn Diagram is the result of one method, as indicated on the top of each Venn Diagram. B. Drug identification results comparison among four signature-based methods with TCGA breast cancer dataset. The Venn Diagrams show the agreement of drug identification results among the four signature-based methods: cMap, sscMap, NFFinder and Cogena. Each Venn Diagram is the result of one sized gene signature, as indicated on the top of each Venn Diagram.

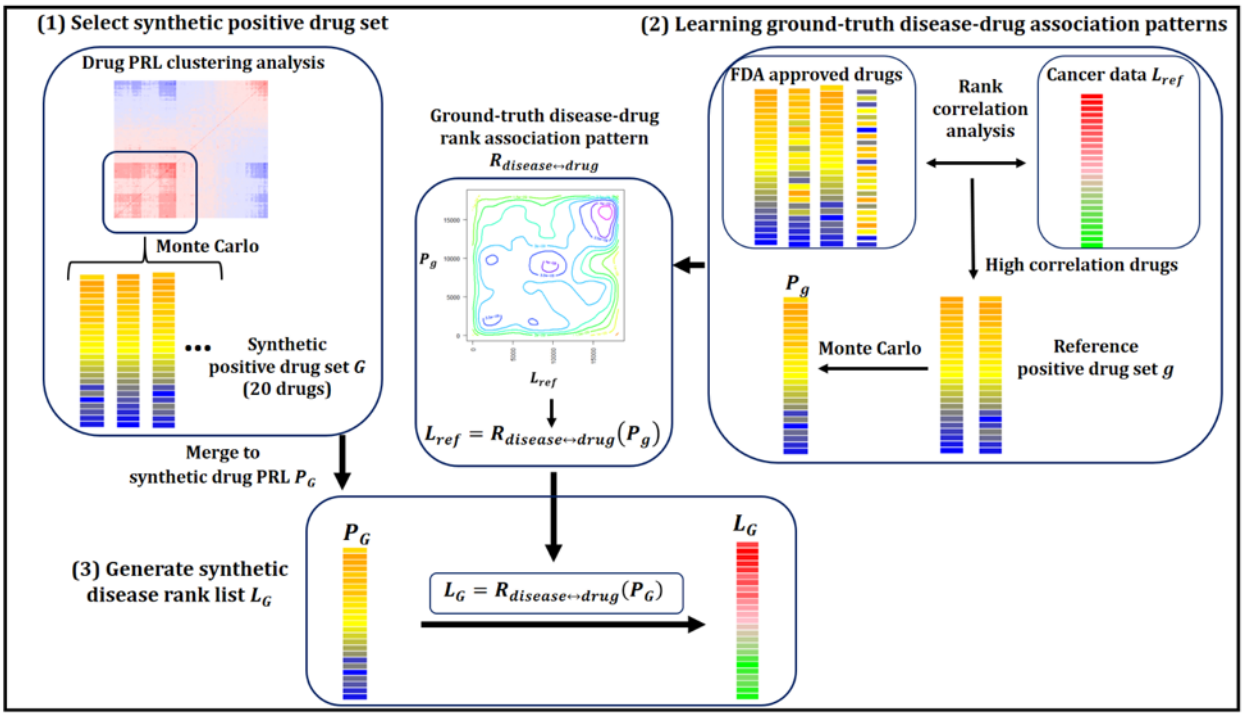

Figure S3. Overview of the simulation process. (1) Select synthetic positive drug set. Based on a highly correlated 697-drug cluster, a positive drug set G is sampled. The merged rank list P_G for G is calculated. (2) Learn ground-truth disease-drug association patterns based on FDA drug set g. The rank correlation pattern $R_{disease\leftrightarrow drug}$ is trained using Monte Carlo sampling. (3) Generate synthetic disease rank list L_G.

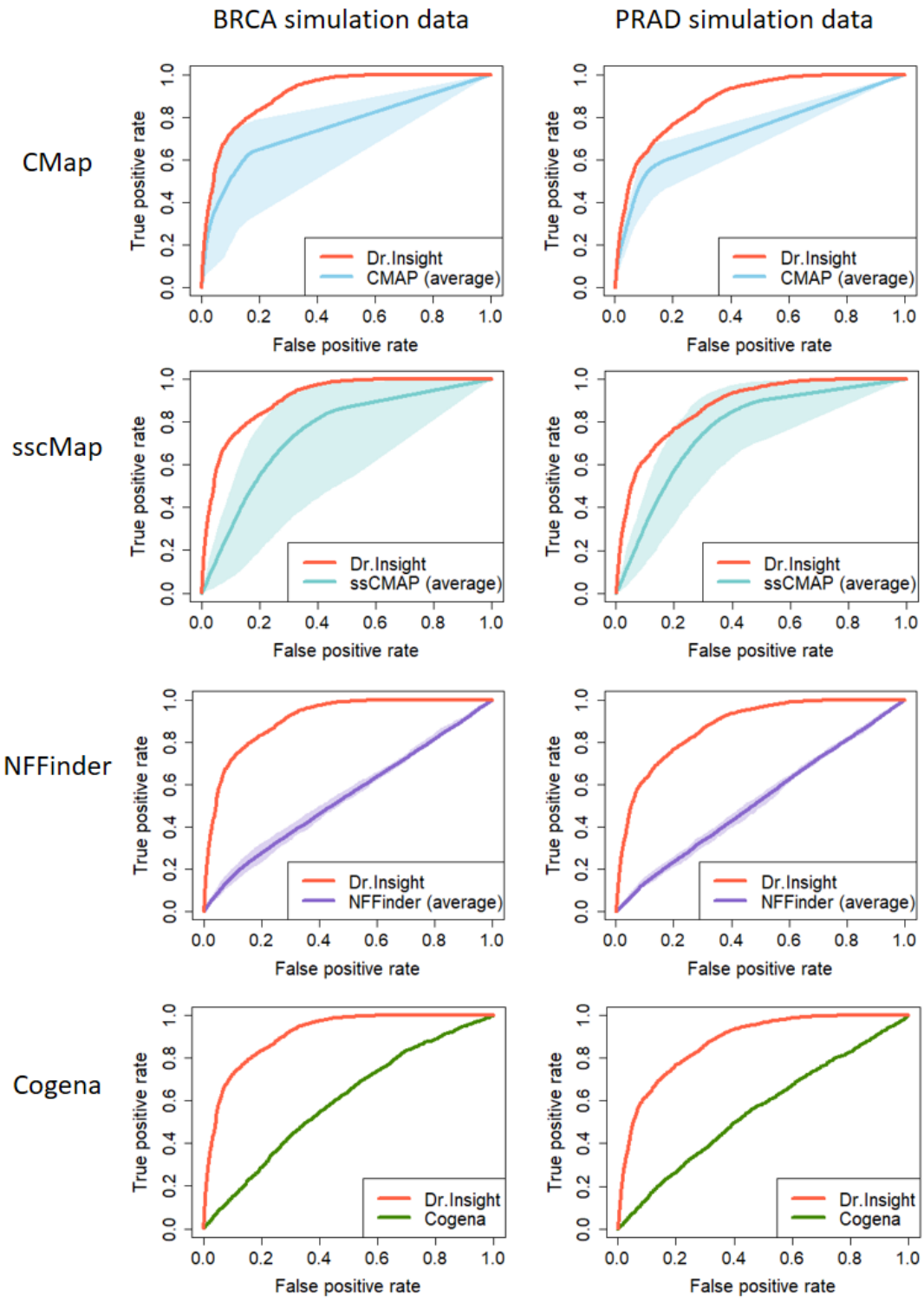

Figure S4. Average ROC curves of Dr. Insight and four signature-based connectivity-mapping methods. For CMap, sscMap and NFFinder, the shaded area represents the range of the ROC curves that corresponds to eight different signature sizes.

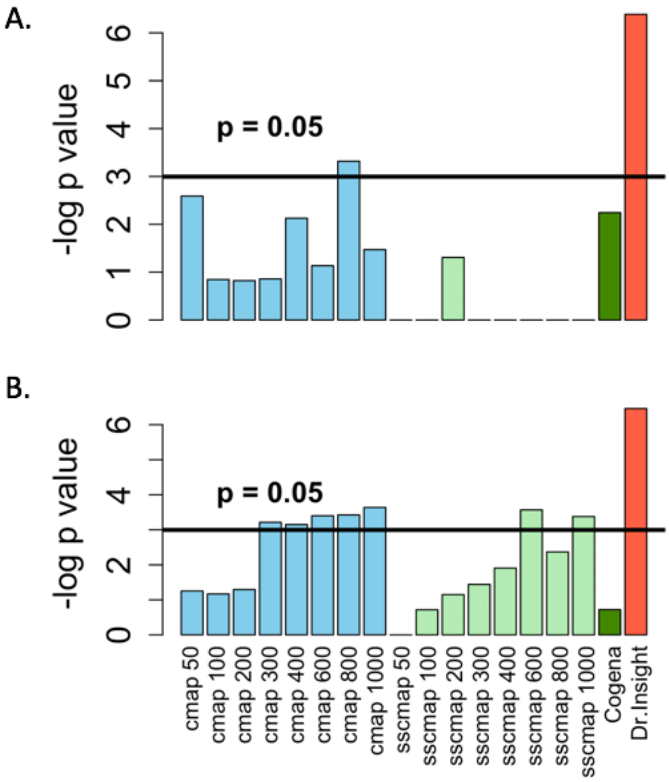

Figure S5. Comparing Dr. Insight with existing methods on non-cancer datasets. The bar plots give the log-transformed enrichment p-values from the four methods. Multiple enrichment p-values are reported for CMap, sscMap, which correspond to query signatures of different sizes. The horizontal lines indicate the 0.05 statistical significance level. A. Systemic lupus erythematosus. B. Hepatitis B virus infection dataset.

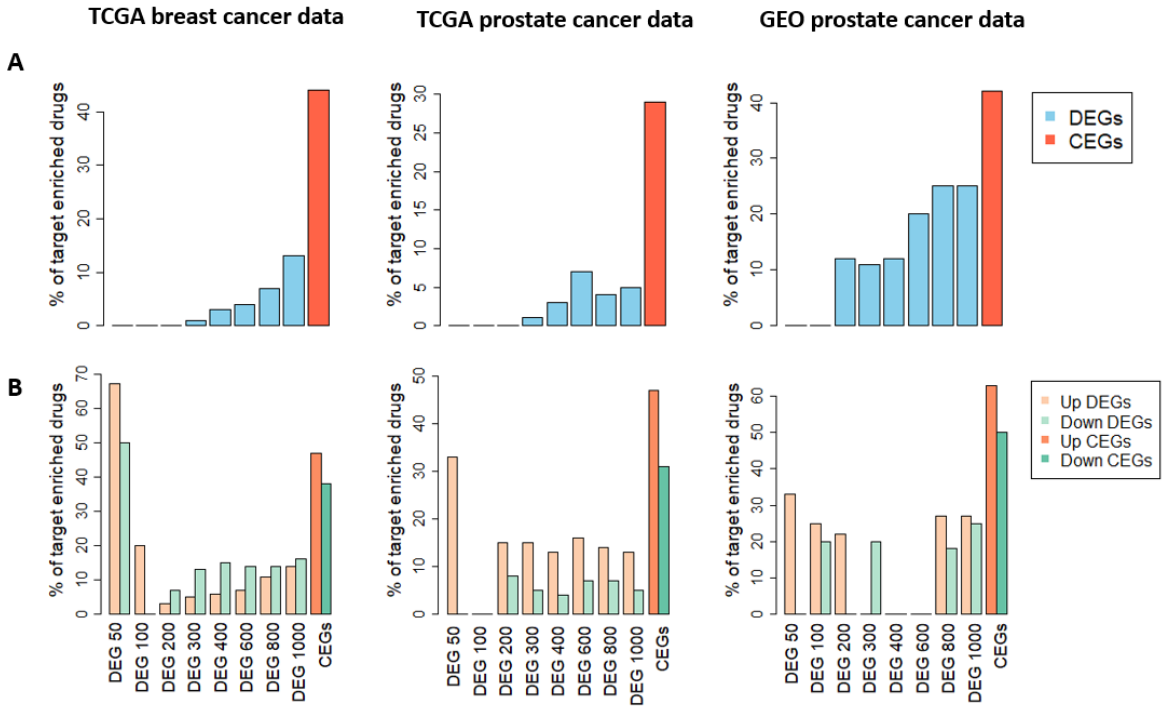

Figure S6. Target enrichment analysis: sscMap vs. Dr. Insight. (A) Results from STITCH database (undirected). (B) Results from CTD database (directed).

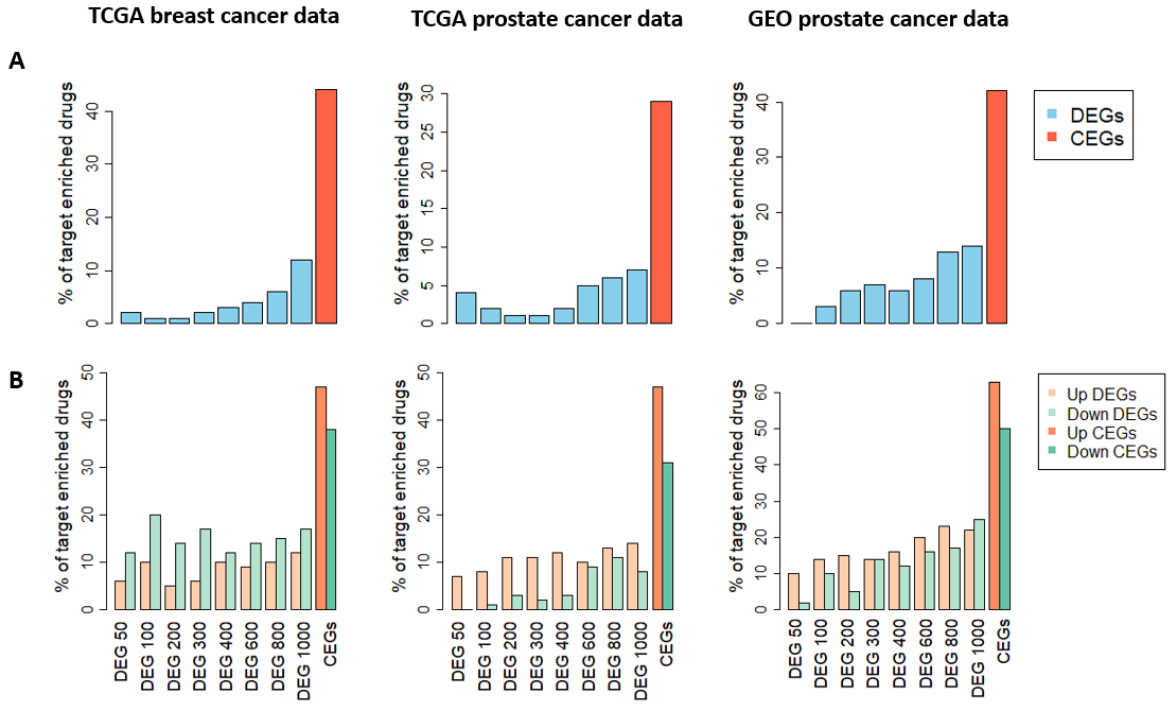

Figure S7. Target enrichment analysis: NFFinder vs. Dr. Insight. (A) Results from STITCH database (undirected). (B) Results from CTD database (directed).

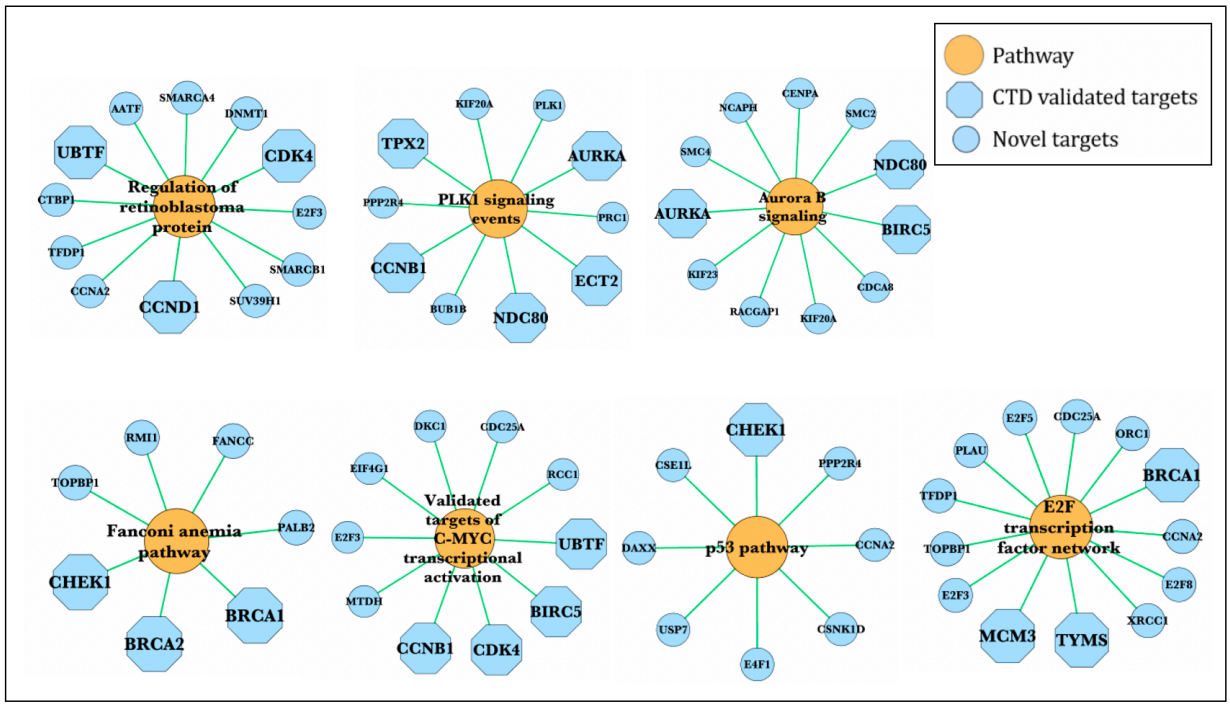

Figure S8. Seven additional target pathways of trichostatin A (TSA) and their potential target genes. Blue hexagons are genes that are identified as CEGs by Dr. Insight, which are also documented as known TSA targets in the CTD database; blue circles are novel, previously undocumented TSA targets.

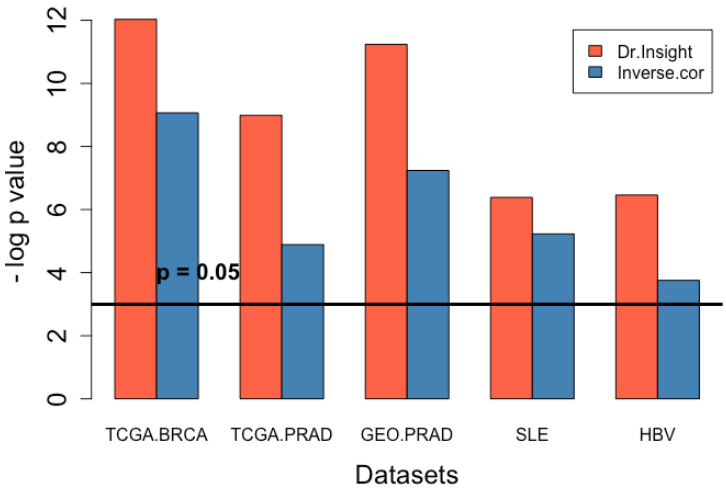

Figure S9. Compare Dr. Insight with the inverse correlation method. The x axis indicates the five datasets: TCGA.BRCA: breast cancer dataset from TCGA. TCGA.PRAD: prostate cancer dataset from TCGA. GEO.PRAD: prostate cancer dataset from GEO. SLE: systemic lupus erythematosus dataset. HBV: Hepatitis B virus infection dataset. The y axis is the log-transformed enrichment p-values from the five methods.

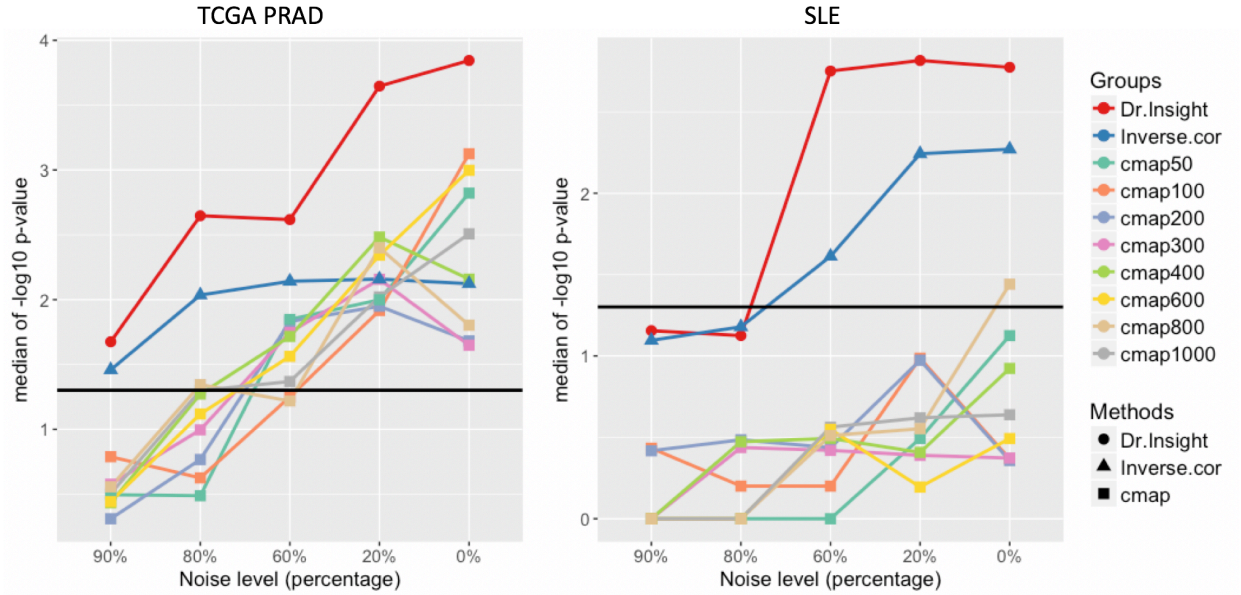

Figure S10. Robustness evaluation against noise. The amount of noise tolerated by three methods has been evaluated using TCGA PRAD (prostate cancer) data and SLE (systemic lupus erythematosus) data. The x axis denotes the percentage of noise added to the original data. The y axis is the median (of 20 repeated experiments) of log-transformed enrichment p-value. Eight different signature sizes were used for CMap.

Abbreviations and Terminologies

CEG: concordantly expressed genes

DEG: differentially expressed genes

ROC: receiver operating characteristic curve

AUC: the area under an ROC curve

BRCA: breast invasive carcinoma

PRAD: prostate adenocarcinoma

SLE: systemic lupus erythematosus

HBV: hepatitis B virus

Drug instance/instance: the gene list ranked by the results of comparison between cells perturbed by drug treatment and cells grown in the same plate and treated with vehicle alone. Each instance represents one replicate of single drug treatment condition (e.g. concentration, cell line, duration, etc.). One drug can have multiple instances where same or different treatment combinations are used.

Drug treatment set: a drug treatment set contains one or multiple drug instances that represent a single drug perturbation on a single cell line, denoted in the analysis results as drug_cellLine, e.g. tamoxifen_MCF7 (S3.2).

References

1. Zhang, S.D. and T.W. Gant, *A simple and robust method for connecting small-molecule drugs using gene-expression signatures.* BMC Bioinformatics, 2008. **9**: p. 258.
2. Varambally, S., et al., *Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression.* Cancer Cell, 2005. **8**(5): p. 393-406.
3. Sinha, S., et al., *Mutant WT1 is associated with DNA hypermethylation of PRC2 targets in AML and responds to EZH2 inhibition.* Blood, 2015. **125**(2): p. 316-26.
4. Borcherding, N., et al., *Paracrine WNT5A Signaling Inhibits Expansion of Tumor-Initiating Cells.* Cancer Res, 2015. **75**(10): p. 1972-82.
5. Chen, H.R., et al., *A network based approach to drug repositioning identifies plausible candidates for breast cancer and prostate cancer.* BMC Med Genomics, 2016. **9**(1): p. 51.
6. Lee, J., et al., *Withaferin A is a leptin sensitizer with strong antidiabetic properties in mice.* Nat Med, 2016. **22**(9): p. 1023-32.
7. Liu, J., et al., *Treatment of obesity with celastrol.* Cell, 2015. **161**(5): p. 999-1011.
8. Sanda, T., et al., *Interconnecting molecular pathways in the pathogenesis and drug sensitivity of T-cell acute lymphoblastic leukemia.* Blood, 2010. **115**(9): p. 1735-45.
9. Peng, G., et al., *Genome-wide transcriptome profiling of homologous recombination DNA repair.* Nat Commun, 2014. **5**: p. 3361.
10. Cheng, H.W., et al., *Identification of thioridazine, an antipsychotic drug, as an antiglioblastoma and anticancer stem cell agent using public gene expression data.* Cell Death Dis, 2015. **6**: p. e1753.
11. Chen, X., et al., *Terazosin activates Pgk1 and Hsp90 to promote stress resistance.* Nat Chem Biol, 2015. **11**(1): p. 19-25.
12. Iorio, F., et al., *Discovery of drug mode of action and drug repositioning from transcriptional responses.* Proc Natl Acad Sci U S A, 2010. **107**(33): p. 14621-6.
13. Nepusz, G.C.a.T., *The igraph software package for complex network research.* InterJournal, 2006. **Complex Systems**: p. 1695.
14. Kleinstein, G.Y.a.C.R.B.a.J.T.a.S.H., *Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations.* Nucleic Acids Res, 2013.
15. Musa, A., et al., *A review of connectivity map and computational approaches in pharmacogenomics.* Brief Bioinform, 2017.
16. Sandmann, T., et al., *gCMAP: user-friendly connectivity mapping with R.* Bioinformatics, 2014. **30**(1): p. 127-8.
17. Huang, H., et al., *DMAP: a connectivity map database to enable identification of novel drug repositioning candidates.* BMC Bioinformatics, 2015. **16 Suppl 13**: p. S4.
18. Pacini, C., et al., *DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data.* Bioinformatics, 2013. **29**(1): p. 132-4.
19. Zhang, S.D. and T.W. Gant, *sscMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures.* BMC Bioinformatics, 2009. **10**: p. 236.
20. O'Reilly, P.G., et al., *QUADrATiC: scalable gene expression connectivity mapping for repurposing FDA-approved therapeutics.* BMC Bioinformatics, 2016. **17**(1): p. 198.
21. Lee, J.H., et al., *CDA: combinatorial drug discovery using transcriptional response modules.* PLoS One, 2012. **7**(8): p. e42573.
22. Shigemizu, D., et al., *Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer.* PLoS Comput Biol, 2012. **8**(2): p. e1002347.

23. Setoain, J., et al., *NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning.* Nucleic Acids Res, 2015. **43**(W1): p. W193-9.
24. Jia, Z., et al., *Cogena, a novel tool for co-expressed gene-set enrichment analysis, applied to drug repositioning and drug mode of action discovery.* BMC Genomics, 2016. **17**: p. 414.
25. Wen, Q., et al., *A gene-signature progression approach to identifying candidate small-molecule cancer therapeutics with connectivity mapping.* BMC Bioinformatics, 2016. **17**(1): p. 211.