bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Network perturbation analysis of gene transcriptional profiles reveals protein targets and mechanism of action of drugs and influenza A viral infection

Heeju Noh[1,2], Jason Shoemaker[3,4] and Rudiyanto Gunawan[1,2,*]

[1] Institute for Chemical and Bioengineering, ETH Zurich, Zurich, 8093, Switzerland
[2] Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland
[3] Department of Chemical and Petroleum Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA
[4] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, 15261, USA

* To whom correspondence should be addressed. Tel: +41 44 633 21 34; Fax: +41 44 633 12 52; Email: rudi.gunawan@chem.ethz.ch

ABSTRACT

Genome-wide transcriptional profiling provides a global view of cellular state and how this state changes under different treatments (e.g. drugs) or conditions (e.g. healthy and diseased). Here, we present ProTINA (Protein Target Inference by Network Analysis), a network perturbation analysis method for inferring protein targets of compounds from gene transcriptional profiles. ProTINA uses a dynamic model of the cell-type specific protein-gene transcriptional regulation to infer network perturbations from steady state and time-series differential gene expression profiles. A candidate protein target is scored based on the gene network's dysregulation, including enhancement and attenuation of transcriptional regulatory activity of the protein on its downstream genes, caused by drug treatments. For benchmark datasets from three drug treatment studies, ProTINA was able to provide highly accurate protein target predictions and to reveal the mechanism of action of compounds with high sensitivity and specificity. Further, an application of ProTINA to gene expression profiles of influenza A viral infection led to new insights of the early events in the infection.

INTRODUCTION

The identification of the molecular targets of pharmacologically relevant compounds is vital for understanding the mechanism of action (MoA) of drugs, as well as for exploring off-target effects. While the definition of a target can be quite arbitrary, the term generally refers to a molecule whose interaction with the compound is connected to the compound's effects (1). In this study, transcription factors (TFs) and their protein interaction partners represent the target molecules, while differential gene expression profiles represent the effects. Among existing technologies for protein target discovery (e.g., biochemical affinity purification, RNAi knockdown or gene knockout experiments) (2), gene expression profiling has received much recent attention due to its relative ease of implementation as well as the availability of large-scale public databases and well-established experimental protocols and data analytical methods. A complication when using gene expression profiling for target discovery is that the data give only indirect indications of the drug's action. As illustrated in Fig. 1a, the interaction between a compound and its protein target(s) is expected to result

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in the differential expression of downstream genes that are regulated by the protein target(s). But, the expression of the protein targets themselves may not – and often do not – change (3). Consequently, target discovery using gene expression profiles requires computational methods to identify the (upstream) targets from the (downstream) effects.

Existing computational strategies for compound target identification using gene expression profiles can generally be classified into two groups: comparative analysis and network-based analysis (4). Comparative analysis methods use the gene expression profiles as drug signatures. Here, the similarity between the differential gene expression of a drug treatment and those of reference compounds or experiments with known targets, implies a closeness in the molecular targets and the MoA (commonly referred to as "guilt by association"). A notable example of such an approach is the Connectivity Map (5), which provides gene expression profiles of human cell lines treated by ~5000 small molecule compounds as queryable signatures for evaluating drug-drug similarities (6). The obvious drawback of comparative analysis methods is their dependence on an extensive and accurate target annotation of the reference gene expression profiles.

In network-based analysis, one adopts a system-oriented view by using cellular networks, such as a gene regulatory network (GRN) and/or a protein-protein interaction network (PIN). A number of network-based analytical methods relied on dynamic models of the GRN to infer network perturbations caused by drug treatment (7-9). Several network-based analytical methods used statistical analysis to score drug targets based on the differential expression of genes that are connected to or regulated by these targets (10-12). Numerous graph-based analyses have also been applied to the gene expression data of drug treatments for target prioritization (3,13,14). More recent methods used a combination of different types of cellular networks. Notably, a method called Detecting Mechanism of Action by Network Dysregulation (DeMAND) combined the GRN and PIN information to create a molecular interaction network, where the drug targets were scored based on drug-induced alterations in the joint gene expression distribution between two connected genes in the network (15). While recent strategies still have some limitations – for example DeMAND could not be used to predict the direction of the drug's effects (e.g. enhancement or attenuation) – the benefit of integrating different biological networks in the analysis of gene expression is clear. As expected, the performance of any network-based analysis would depend on the fidelity of the underlying network.

In this work, we developed ProTINA (Protein Target Inference by Network Analysis), a network perturbation analysis method for protein target identification using gene transcriptional profiles. The analysis involves two key steps: (a) the creation of a model of tissue or cell type-specific protein-gene regulatory network (PGRN), and (b) the calculation of protein target scores based on the enhancement or attenuation of the protein-gene regulations. ProTINA leverages on the availability of comprehensive maps of protein-protein and protein-DNA interactions for the construction of the PGRN graph, and employs ridge regression to infer the protein-gene regulatory activity from differential gene expression profiles. We demonstrated the superiority of ProTINA over the state-of-the-art method DeMAND and differential gene expression analysis (DE), in predicting the protein targets of drugs. Besides protein targets of compounds, we presented the application of ProTINA to study host-pathogen interactions, specifically for elucidating the targets of influenza A viral proteins.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MATERIAL AND METHODS

Gene expression data

We applied ProTINA to three datasets of drug treatments from NCI-DREAM drug synergy challenge (16), genotoxicity study (17) and chromosome drug targeting study (18), and to gene expression data of human lung cancer cell Calu-3 from influenza A viral infection studies (19-22). For NCI-DREAM drug synergy challenge, we obtained the raw *Affymetrix Human Genome U219* microarray data from Gene Expression Omnibus (GEO) database (23) (accession number: GSE51068). The raw data were first normalized and transformed into log2-scaled expressions using *justRMA* function in the *affy* package of Bioconductor (24). Then, the $\log_2$ fold change (log2FC) differential expressions and their statistical significance (Benjamini-Hochberg adjusted *p*-values) were calculated using a linear fit model and empirical Bayes method in the *limma* package of Bionconductor. Three samples from the drug treatment using the low dose of Aclacinomycin A were dropped because all of the log2FC expressions were close to 1 and thus not statistically significant. The probe sets were mapped to gene symbols using *hgu219.db* annotation package (Entrez Gene database as of 27th September 2015). In the case of multiple probe sets mapping to a gene symbol, we assigned the log2FC from the probe set with the smallest average adjusted *p*-value over the samples.

The raw microarray data from genotoxicity study (17) in human HepG2 cell line were obtained from GEO (accession numbers: GSE28878 using *Affymetrix GeneChip Human Genome U133 Plus 2.0* array and GSE58235 using *Affymetrix HT Human Genome U133+ PM* array). As with the drug synergy data, the microarray data were first normalized using *justRMA*, and the log2FCs and their adjusted *p*-values were calculated using *limma* in Bioconductor. Because the data came from different microarray platforms, the gene symbols were matched separately for each platform using *hgu133plus2.db* annotation package (Entrez database of 27th September 2015) and *HT_HG-U133_Plus_PM* annotation file in Affymetrix, respectively. Likewise, in the case of multiple probe sets matching a gene symbol, the probe set with the smallest average adjusted *p*-value across all samples was chosen.

The raw data from the chromosome-targeting study using mouse pancreatic alpha and beta cells (18) were also obtained from GEO database (ascension number: GSE36379). Again, the raw data were normalized using *justRMA*, and the log2FCs and their adjusted *p*-values were calculated by *limma*. The probes were mapped to the corresponding gene symbols using *moe430a.db* package (Entrez Gene database as of 27th September 2015) in Bioconductor. In the case of multiple probe sets mapping to a gene symbol, we selected the probe set with the smallest average adjusted *p*-value among the samples.

For influenza A infection analysis, we obtained the raw microarray data of four influenza studies (19-22) from GEO database (ascension numbers: GSE40844, GSE37571, GSE33142, andGSE28166). The raw data were background-corrected and normalized using *normexp* and *quantile* methods in *limma* package of Bioconductor. The log2FCs and their adjusted *p*-values were again calculated by *limma*. The probes were mapped to the corresponding gene symbols using hgug4112a.db package (Entrez Gene database as of 27th September 2015). Like before, for genes

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with multiple probe sets, we chose the logFC value corresponding to the probe set with the smallest average adjusted *p*-value.

Protein target identification using ProTINA

Protein-gene network. In ProTINA, the PGRN is a bipartite graph with weighted, directed edges pointing from a protein to a gene (see **Fig. 1a**). The edges in the PGRN describe the regulation of gene expression by TFs and their protein partners. The PGRN is constructed by combining two types of networks, namely the TF-gene network and PIN. For the construction of human cell type-specific PGRNs, we relied on the Regulatory Circuit resource that provides 394 cell type and tissue-specific TF-gene interactions (25). More specifically, for the analysis of the NCI-DREAM drug synergy, genotoxic compound study, and influenza A viral infection study data sets, we used the TF-gene networks of human lymphoma cells, pleomorphic hepatocellular carcinoma cells, and epithelium lung cancer cells, respectively. We included only TF-gene interactions with a Regulatory Circuit confidence score greater than 0.1. The confidence score indicates the normalized promoter activity level in a given cell type (0: not active, 1: maximally active) [25]. For the analysis of mouse pancreatic cell dataset, we obtained the mouse pancreatic TF-gene interactions from CellNet (26). In the construction of the PGRNs, any TF-gene interactions involving unmeasured genes were excluded. In summary, the TF-gene network for human lymphoma, hepatocellular carcinoma cell, and epithelium lung cancer cell lines included 31,392 edges pointing from 515 TFs to 5,153 genes, 3,868 edges pointing from 413 TFs to 953 genes, and 42,145 edges pointing from 515 TFs to 7,125 genes, respectively. The mouse pancreatic PGRN included 2,922 edges, involving 95 TFs and 588 genes.

For human PIN, we combined the protein-protein interactions from two databases, namely Enrichr (27) and STRING (28). For mouse pancreatic cells, we obtained mouse (*Mus musculus*) PIN from the STRING database (28). For each TF, we identified its protein partners, defined as proteins that are within a network distance of 2 from the TF in the PIN. When using the STRING database, we included all direct protein partners of TFs, and proteins with a network distance of 2 from TFs with a confidence score reported on STRING larger than 0.5. For human lymphoma, hepatocytes, and lung cancer cells, we identified 11,090 protein partners for a subset of 499 TFs (out of 515 TFs), 10,834 protein partners for a subset of 403 TFs (out of 413 TFs), and 6,175 protein partners for a subset of 504TFs (out of 515 TFs), respectively. For mouse pancreatic cells, we found 6,620 protein partners for a subset of 89 TFs (out of 95 TFs).

Finally, in the construction of the PGRNs, we assigned a directed edge from a TF or from a protein partner of a TF, to every gene regulated by the TF. In summary, the cell type-specific PGRN for human lymphoma cells included 21,488,617 regulatory edges among 11,161 TFs/proteins and 5,153 genes. For hepatocellular carcinoma cells, the PGRN comprised 3,726,393 edges among 10,893 TFs/proteins and 953 genes. For human lung cancer cells, the PGRN comprised 30,656,861 edges among 11346 TFs/proteins and 7,125 genes. For mouse pancreatic cells, the PGRN consisted of 1,417,972 edges among 6,661 TFs/proteins and 588 genes. While increasing the size of the PGRN, for example by including lesser confident TF-gene and protein-protein interactions or by including

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

proteins with a network distance from TFs larger than 2, would allow the scoring of a higher number of proteins, such strategy often lowers the accuracy of the protein target predictions.

Gene transcription model. The edges in the PGRN have weights, whose magnitudes represent the strength of the gene regulation and whose signs indicate the direction or the mode of the regulation: positive for gene activation and negative for gene repression. The weights are inferred from the gene expression dataset by adapting a procedure described in our previous method DeltaNet (9,29). The inference of the edge weights is based on an ordinary differential equation (ODE) model of the mRNA production of a gene:

$$\frac{d\,r_k(t)}{dt} = u_k \prod_{j=1}^{n} r_j(t)^{a_{kj}} - d_k\, r_k(t) \tag{1}$$

where $r_k(t)$ is the mRNA concentration of gene k at time t, u_k and d_k denotes the mRNA transcription and degradation rate constants respectively, and a_{kj} denotes the gene regulatory influence (or edge weight) of the j-th protein on the k-th gene.

While the regulatory edges in the model above usually describe TF-gene interactions, in ProTINA, we further accounted for the (indirect) regulation of a gene by proteins that interact with the TFs. For this purpose, we considered a modified ODE model:

$$\frac{d\,r_k}{dt} = u_k \left(\prod_{j=1}^{n_{TF}} r_j^{\,a_{kj}} \prod_{q=1}^{n_P} (r_j\, r_q)^{b_{kjq}} \right) - d_k\, r_k \tag{2}$$

where a positive (negative) b_{kjq} describes the activation (repression) of the k-th gene by a protein q through its interaction with the TF protein j. The variables n_{TF} and n_P denote the numbers of TFs and their protein partners, respectively. The multiplication of two variables r_j and r_q implies that the regulation of gene k by protein q requires the TF protein j (a non-zero r_j). The model in Equation (2) can be simplified into:

$$\begin{aligned} \frac{d\,r_k}{dt} &= u_k \left(\prod_{j=1}^{n_{TF}} r_j^{\,a_{kj} + \sum_q b_{kjq}} \right) \left(\prod_{q=1}^{n_P} r_q^{\,\sum_j b_{kjq}} \right) - d_k r_k \\ &= u_k \left(\prod_{j=1}^{n_{TF}} r_j^{\,a^*_{kj}} \right) \left(\prod_{q=1}^{n_P} r_q^{\,a^*_{kq}} \right) - d_k r_k \\ &= u_k \left(\prod_{j=1}^{n_{TF}+n_P} r_j^{\,a^*_{kj}} \right) - d_k r_k \end{aligned} \tag{3}$$

where a^*_{kj} denotes the overall regulatory influence of each protein j, including TFs and their protein partners, on the expression of gene k. Note that the model in Equation (3) is mathematically equivalent to that in Equation (1).

By taking the pseudo-steady state assumption, the above model equation can be linearized using a logarithmic transformation (see derivation in ref. 9). The inference of the weights from the gene expression dataset involved the following linear regression problem:

$$c_{ki} = \sum_{j=1}^{n_{TF}+n_P} a^*_{kj} c_{ji} + p_{ki} \tag{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where c_{ki} denotes the log2 fold-change (log2FC) expression for gene k in sample i. The variable p_{ki} represents the part of log2FC of gene k expression in sample i that cannot be accounted for by the log2FC of its protein regulators. In other words, p_{ki} indicates the perturbations to the expression of gene k. As detailed below, ProTINA relies on the magnitude and directions of such network perturbations (dysregulations) to identify proteins with altered gene regulatory activity.

The dynamical information contained in time-series gene expression profiles could greatly improve the inference of the edge weights above. As previously described in ref. 29, such information could be accounted for by adding the following linear constraint on the linear regression problem:

$$s_{ki} = \sum_{j=1}^{n_{TF}+n_P} a_{kj}^{*} s_{ji} \tag{5}$$

where s_{ki} is the time derivatives (slope) of the log2FC of gene k in sample i. The slopes of the log2FC at each sampling time point were computed using a second-order accurate finite difference approximation (30). In summary, the estimation of edge weights in ProTINA involved the following linear regression problem:

$$\mathbf{C}_k = \mathbf{A}_k \mathbf{C}_{R_k} + \mathbf{P}_k \tag{6}$$

$$\mathbf{S}_k = \mathbf{A}_k \mathbf{S}_{R_k} \tag{7}$$

where $\mathbf{C}_k$ and $\mathbf{S}_k$ are the $1\times m$ vectors of log2FC expressions and time-derivatives of gene k across m samples, the subscript R_k refers to the set of ($n_{TF,k}$+$n_{P,k}$) protein regulators of gene k in the cell type-specific PGRN, $\mathbf{C}_{R_k}$ and $\mathbf{S}_{R_k}$ denote the (n_{TF}+$n_{P,k}$)$\times m$ matrices of log2FCs and their slopes across m samples, $\mathbf{A}_k$ is the $1\times(n_{TF}+n_P)$ vector of weights for edges in the PGRN pointing to gene k, and $\mathbf{P}_k$ is the $1\times m$ vector of dysregulation impacts of gene k over m samples.

In ProTINA, the vectors $\mathbf{A}_k$ and $\mathbf{P}_k$ for each gene k in Equations (6) and (7) were estimated by ridge regression. The ridge regression provides a solution to an underdetermined linear regression problem of the standard form: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, using a penalized least square objective function:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

where λ is a shrinkage parameter for the L^2-norm penalty. Equations (6) and (7) are rewritten into the standard linear regression problem with $\mathbf{y} = [\mathbf{C}_k \;\; \mathbf{S}_k]^{\mathrm{T}}$, $\mathbf{X} = [\,[\,\mathbf{C}_{R_k} \;\; \mathbf{S}_{R_k}\,]^{\mathrm{T}},\; [I_m \;\; \mathbf{0}]^{\mathrm{T}}\,]$, $\beta = [\mathbf{A}_k \;\; \mathbf{P}_k]^{\mathrm{T}}$. Before applying the ridge regression, we normalized the vectors of log2FCs and slopes to have a unit norm. Self-loops were excluded in the regression, and thus the diagonal entries of $\mathbf{A}_k$ were set to 0. In the applications of ProTINA, we employed 10-fold cross validations to determine the optimal λ, one that gives the minimum average prediction error. Here, we used the GLMNET package (31) for both the MATLAB and R versions of ProTINA.

Protein target scoring. In ProTINA, each candidate protein target is assigned a score based on the deviation of the expression of its downstream genes. More specifically, we computed the residuals of the linear regression problem in Equations (6) for each gene k, i.e.

$$\mathbf{r}_k = \mathbf{C}_k - \mathbf{A}_k \mathbf{C}_{R_k} \tag{9}$$

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\mathbf{r}_k$ is the $1\times m$ vector of residuals for m samples. For each drug treatment, there often exist multiple gene expression profiles, taken at different time points or different doses. Correspondingly, we evaluated the z-score z_{lk} for each drug treatment l and for each gene k, according to

$$z_{lk} = \frac{\overline{r}_{lk}}{\sigma_k / \sqrt{n_l}} \tag{9}$$

where $\overline{r}_{lk}$ denotes the average residual of gene k among the drug treatment samples, σ_k denotes the sample standard deviation of the residuals in all samples besides the drug treatment, and n_l denotes the number of samples from the drug treatment. A positive (negative) z-score indicates that the expression of gene k in the particular sample was higher (lower) than expected based on the expression of its regulators. The greater the magnitude of the z-score, the more significant is the gene dysregulation.

The target score of a TF or protein for a drug is calculated by combining the z-scores of the target genes in the PGRN, as follows: (ref. 32)

$$s_{ji} = \frac{\sum_{k=1}^{n_D} w_{kj} z_{ki}}{\sqrt{\sum_{k=1}^{n_D} w_k^2}} \tag{10}$$

where z_{ki} denotes the z-score of gene k and s_{ji} denotes the score of the TF/protein j in the drug treatment sample i. The weighting coefficients w_{kj} are set equal to the edge weights a_{kj} divided by the maximum magnitude of a_{kj} across all j. In other words, the weight w_{kj} reflects the fraction of the regulation of gene k expression that could be attributed to protein j. When w_{kj} (or a_{kj}) and z_{ki} have the same signs, $w_{kj}z_{ki}$ thus takes a positive value. As illustrated in **Fig. 1c**, a positive $w_{kj}z_{ki}$ implies an enhanced regulatory activity of protein j on gene k, since the activation (inhibition) of gene k expression by protein j is stronger in this sample than expected by the PGRN model. In contrast, a negative $w_{kj}z_{ki}$ indicates an attenuation of the regulatory influence of protein j on gene k, since the activation (inhibition) of gene k expression by protein j is weaker than predicted by the PGRN model. Consequently, a highly positive (negative) score s_{ji} is an overall indicator of strongly enhanced (attenuated) regulatory activity of protein j by the drug treatment in sample i (see **Fig. 1d**). The protein targets in each drug treatment sample are ranked in decreasing magnitude of the scores s_{ji}.

DeMAND and differential expression analysis

For DeMAND analysis, we employed the public R subroutines available from the website: http://califano.c2b2.columbia.edu/demand. Following the procedure detailed in the original publication (15), we computed the RMA (Robust Multi-array Average) normalized gene expression values as inputs to the analysis. In DeMAND analysis, we used the same cell type-specific PGRNs as those in ProTINA. For each candidate protein target, DeMAND evaluated the *p*-value of the deviations in the gene expression relationship between the protein target and each of the genes connected to this protein in the PGRN. The drug targets were ranked in increasing magnitude of the combined *p*-values.

In differential expression (DE) analysis, we calculated the log2FC differential expression of each protein in the PGRN, as described in section **Gene expression data** above. Here, we used the

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

log2FC values directly as the target scores. Correspondingly, we ranked the candidate protein targets in decreasing magnitude of the log2FC gene expression values.

Performance assessment

For comparing the performance of different methods, we computed the area under the receiver operating characteristic curve (AUROC), i.e. the area under the plot of true positive rate against false positive rate, following the procedure adopted in DREAM challenges (33,34). For each method and each drug treatment, we generated a ranked list of protein targets according to decreasing magnitudes of the protein scores in ProTINA, increasing *p*-values of network dysregulation from DeMAND, and increasing magnitudes of log2FC gene expression from DE analysis.

Gene set enrichment analysis

For influenza A virus study, we performed a gene set enrichment analysis (GSEA) of the protein target predictions from ProTINA, DeMAND and DE analysis for the KEGG biological pathways (35), using the R package GAGE (Generally Applicable Gene-set/pathway Enrichment analysis) with Kolmogorov-Smirnov tests (36). In the case of ProTINA and DeMAND, target proteins with zero score were excluded from the GSEA.

Reference protein targets

The reference protein targets of compounds in drug treatment studies were compiled from 5 different public databases of chemical-protein interactions: DrugBank (37), Therapeutic Target Database (TTD) (38), MATADOR (39), Comparative Toxicogenomics Database (CTD) (40), and STITCH (41). DrugBank and TTD provided information on the mechanism of drug actions as well as the proteins that have physical binding interactions with drugs. Meanwhile, MATADOR, CTD, and STITCH gave interactions between proteins and chemical compounds, curated from text mining and experimental evidences. When retrieving the protein targets of drugs from these databases, we collected proteins that directly bind to the queried drugs. The reference targets for each dataset in this study are provided in **Supplementary material 1**. Meanwhile, the reference protein targets for influenza A virus study were obtained from ref.42, where 1,292 host proteins that likely physically bind to viral proteins of influenza type A/WSN/33 in human embryonic kidney cells (HEK293) were identified by whole-genome co-immunoprecipitation assays.

RESULTS

New protein target prediction strategy

ProTINA takes advantage of the availability of comprehensive protein-protein and protein-DNA interaction databases to construct, when possible, a tissue or cell type-specific PGRN. The method considers a PGRN with weighted directed edges (see **Fig. 1a**), describing direct and indirect gene

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transcriptional regulation by TFs and their protein partners. The edge weights are determined by applying ridge regression using the gene expression data based on a kinetic model of the gene transcriptional process (see **Fig. 1b** and **Material and Methods**). Here, a positive weight indicates a gene activation, while a negative weight implies a gene repression. Because of the underlying kinetic model, ProTINA is able to incorporate dynamical gene expression data, a common type of data from drug treatment studies (5,16-18). The scoring of drug targets is based on the enhancement or attenuation of protein-gene regulatory interactions caused by the drug treatment. A drug-induced gene regulatory enhancement occurs when the expression of genes that are positively (negatively) regulated by a candidate target, becomes higher (lower) in drug treated samples than what is predicted by the PGRN model (see **Fig. 1c**). A drug-induced attenuation describes the opposite scenario, where the expression of positively (negatively) regulated genes of a target is lower (higher) than expected from the model. For any given differential gene expression sample, a candidate protein target is scored based on the overall enhancement and/or attenuation of its regulatory influence on the downstream genes (see **Fig. 1d** and **Material and Methods**). Thus, a protein target with a more positive (negative) score is considered a more likely target of the drug, in which the drug treatment enhances (attenuates) the gene regulatory activity.

Prediction of known targets of drugs

We tested ProTINA's performance in predicting drug targets using gene expression data from three drug treatment studies employing human and mouse cell lines. The first dataset came from the NCI-DREAM drug synergy study using human diffuse large B cell lymphoma OCI-LY3 (16), the second from the compound genotoxicity study using human liver cancer cells HepG2 (17), and the third from the chromatin-targeting compound study using mouse pancreatic cells (18). We compared ProTINA to the state-of-the-art network-based analytical method DeMAND (15), and to the traditional differential expression analysis (DE). For the analysis of datasets from human cell lines, we constructed cell-type specific PGRNs by combining human PIN from STRING (28) and Enrichr database (27) and human cell-type specific protein-DNA networks from Regulatory Circuit resource (25). Meanwhile, for the construction of mouse pancreatic cell type-specific PGRN, we used mouse (*Mus musculus*) PIN from STRING (28) and mouse protein-DNA interactions from CellNet (26) (see details in **Material and Methods**).

In assessing the performance of ProTINA and the other methods, we compared the ranked list of protein target predictions for each compound with the reference drug targets compiled from the literature (see **Material and Methods** and **Supplementary material 1**). **Fig. 2** (also see **Supplementary Table S1-3**) summarizes the AUROCs of the target predictions from ProTINA, DeMAND, and DE analysis, showing ProTINA significantly outperforming DeMAND and DE analysis for all three datasets. Here, the drug target predictions from DE analysis had the poorest AUROCs with an overall average below 0.664 (AUROC range: 0.393 – 0.982). Meanwhile, the target predictions of DeMAND were slightly better than the DE analysis, averaging at 0.743 (AUROC range: 0.405 – 0.989) for the three datasets. Meanwhile, ProTINA gave the highest average AUROCs among the methods with an average of 0.825 (AUROC range: 0.425 – 0.991).

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mechanism of action of drugs

Besides high AUROCs, ProTINA also provided accurate and specific indications on the MoA of the compounds. In the NCI-DREAM synergy study, roughly half of the compounds are known to cause DNA damages, including DNA topoisomerase inhibitors (camptothecin, doxorubicin and etoposide), DNA crosslinker (mitomycin C), oxidative DNA damaging agent (methothrexate), and histone deacetylase (HDAC) inhibitors (trichostatin A). In demonstrating ProTINA's ability to reveal the compound MoA, we focused on the canonical p53 DNA damage response pathway (15), as illustrated in **Fig. 3**. Here, the activation of p53 in response to DNA damage is expected to induce the transcription of Cyclin Dependent Kinase Inhibitor 1A (CDKN1A) and Growth Arrest and DNA Damage Inducible Alpha (GADD45A) (43,44). In turn, CDKN1A and GADD45A – through their interactions with Proliferating Cell Nuclear Antigen (PCNA) – regulate the DNA replication and repair process (45). GADD45A also inhibits the catalytic activity of Aurora Kinase A (AURKA) (46), leading to a lowered activation of Polo-like Kinase 1 (PLK1) and Cyclin B1 (CCNB1) in a phosphorylation cascade (47,48). As shown in **Fig. 4a**, except for trichostatin A, the six proteins in the canonical p53 pathway above were ranked highly by ProTINA among the genotoxic compounds in the study (median rank <500), consistent with their known MoA. Note that the same six proteins were ranked much lower among the non-DNA damaging compounds (median rank >500), signifying a high specificity of ProTINA predictions (see also Supplementary Figure S1). Equally important, ProTINA was able to accurately identify the direction of the drug-induced alterations caused by the DNA damaging compounds. The signs of protein target scores from ProTINA indicated drug-induced enhancement (positive scores) of CDKN1A, PCNA, and GADD45A, and attenuation (negative scores) of CCNB1, AURKA, and PLK1 (see **Supplementary Table S4**), consistent with the expected response of these proteins to DNA damage in **Fig. 3**.

As illustrated in **Fig. 4a**, DeMAND and DE analysis also performed reasonably well in predicting the compounds' MoA. But, the directions of the perturbations predicted by DE analysis were not always consistent with the expected response to DNA damage (see **Supplementary Table S5-6**). Meanwhile, DeMAND did not provide any information on the directions of the drug perturbations. In addition, the protein target scores of ProtTINA provided a clearer demarcation between the genotoxic and the non-genotoxic agents among the compounds in the dataset, than DeMAND and DE analysis (see Supplementary Figure S1). Besides the canonical p53 response pathway, we further looked at the ranking of proteins involved in the overall DNA damage repair (DDR) and its associated pathways (49) (see **Supplementary material 2**). As depicted in **Fig. 4b**, ProTINA ranked these proteins much higher than DeMAND and DE analysis, with DE performing the poorest among the methods considered.

In comparison to DeMAND and DE analysis, ProTINA was further able to detect a specific MoA of mitomycin C, whose DNA crosslinking activity is expected to prompt a particular DNA repair process called the fanconi anemia pathway (50). The fanconi anemia pathway relies on a specific protein complex to ubiquitinate Fanconi Anemia Group D2 Protein (FANCD2) and Fanconi Anemia Group I Protein (FANCI), as well as two homologous recombination (HR) repair proteins, namely Breast

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cancer Type 1 Susceptibility Protein (BRCA1) and RAD51 Recombinase (RAD51) (51). In ProTINA analysis, the average rank of FANCD2, FANCI, BRCA1, and RAD51 was within top 100 for mitomycin C, while the average rank of those proteins was much greater than 100 for the other DNA damaging agents (see **Supplementary Table S7**). However, the specific activation of the fanconi anemia pathway by mitomycin C was not detected by DeMAND or DE analysis. Thus, ProTINA provided more sensitive and specific indications for the mechanism of action of compounds than DeMAND and DE.

Application of ProTINA for predicting pathogen-host interactions

We applied ProTINA to time-course gene expression profiles of human lung cancer cells (Calu-3) under influenza A virus infection, with the goal of identifying host factors that interact with the viral proteins. The gene expression data came from four studies of influenza A viruses, including A/Netherlands/602/2009 (H1N1), A/CA/04/2009 (H1N1), and A/Vietnam/1203/2004 (H5N1) (19-22). We employed ProTINA to compute the overall protein target scores using the gene expression data of Calu-3 from the four studies above, by averaging the scores from the early phase of the influenza infection between 0 to 12 hours. We checked the target predictions of ProTINA against the findings from a genome-wide co-immunoprecipitation analysis of host and viral protein interactions (42). More specifically, the aforementioned study reported 1,292 host proteins that co-immunoprecipitated with viral proteins of influenza A/WSN/33 using human embryonic kidney cells (HEK293). Despite the discrepancy in the cell types and influenza viral strains between the co-immunoprecipitation analysis and the gene expression profiling, influenza A viruses share similar features and common protein interactions (52,53). Besides ProTINA, we also evaluated the accuracy of viral target predictions from DeMAND and DE for the same dataset.

Fig. 5 gives the receiver operating characteristic (ROC) curves of the target predictions from ProTINA, DeMAND and DE analysis. ProTINA outperformed the two other methods, providing the highest AUROC (ProTINA: 0.758 vs. Demand: 0.687 and DE: 0.647). We further performed a gene set enrichment analysis (GSEA) for the target predictions from each of the methods (see **Material and Methods**) to elucidate the key pathways involved in the viral infection and the accompanying host response. The results of the GSEA are summarized in **Fig. 6**. Both DeMAND and DE target predictions were enriched for only a few pathways (q-value < 0.01), while ProTINA prediction had a much higher number of overrepresented pathways.

The common enriched pathways among ProTINA, DeMAND and DE (top of **Fig. 6**) included known mechanisms related to viral entry, replication and assembly, including endocytosis (54), protein processing in endoplasmic reticulum (55), ubiquitin mediated proteolysis (56,57) and RIG-I-like receptor signaling pathway (58,59). Both ProTINA and DE analysis indicated the modulation of host cell cycle (60), mRNA surveillance (61) and DNA damage response (62). Only ProTINA prediction was significantly enriched for focal adhesion and actin cytoskeleton, which have been shown to regulate influenza virus entry at the early stage of infection (63). In addition, ProTINA target predictions were also enriched for a broad set of host response pathways to viral infection, including host defense mechanism (e.g., T- and B-cell receptor pathways, phagocytosis, leukocyte migration, chemokine signaling pathways), DNA damage repair (e.g., nucleotide excision repair, p53 signaling

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pathway, homologous recombination) and apoptosis. As several influenza proteins are known to interfere with interferon production (which in turn regulates several cytokines) (58,59), these findings suggest that, overall, ProTINA provided a broader picture of the early events in the influenza A viral infection, than DeMAND and DE analysis.

DISCUSSION

ProTINA is a novel and highly effective network-based analytical method for inferring the protein targets of compounds from gene expression profiling data. The target scoring is based on quantifying perturbations in the protein-gene regulatory network, specifically enhancement or attenuation of gene regulatory interactions, caused by the compound treatment. ProTINA combines information of TF-gene and protein-protein interactions and data of differential gene expressions to create a tissue or cell type-specific PGRN model. In the applications to three benchmark drug treatment datasets using human and mouse cell lines, ProTINA significantly outperformed the state-of-the-art algorithm DeMAND, which also relies on network dysregulation scoring, and the standard DE analysis. The target predictions of ProTINA also provide indications for the MoA of compounds, including the directions of the network perturbations, with high sensitivity and specificity.

Both ProTINA and DeMAND score the protein targets of compounds based on gene regulatory network perturbations. In particular, DeMAND calculates protein dysregulation scores (p-values) for a given gene regulatory network, by statistical comparison between samples from drug treatment and from control experiments. Thus, DeMAND requires only few samples to generate its prediction (provided that the network can be defined *a priori*). On the other hand, ProTINA makes use of available differential gene expression profiles from a study or a cell line (i.e., not only for a particular drug), to assign the edge weights of the PGRN by ridge regression. Importantly, in the regression analysis, the PGRN model used in ProTINA accounts for the network perturbations. The ability of ProTINA to incorporate data from other drug treatments or conditions in the scoring of protein targets makes this method particularly suited to take advantage of extensive and still growing number of gene transcriptional profiles from publicly accessible databases such as GEO.

Besides its intended use to predict targets of compounds, we also demonstrated that the analysis of network perturbations using ProTINA could provide insights into the mechanism of diseases. In the application to gene expression profiles of Calu-3 cells from influenza A infection studies, ProTINA again outperformed DeMAND and DE analysis in identifying host factors that bind with viral proteins. Furthermore, the GSEA of ProTINA target predictions revealed the spectrum of cellular processes involved in the early phase of influenza A infection, including pathways involved in viral entry, replication and assembly, and those related to cellular response to viral infection. Among the pathways with the highest significance (lowest *q*-value) was focal adhesion, which has been shown to regulate influenza viral entry as well as replication (63). Meanwhile, the target predictions of DeMAND and DE analysis had fewer enriched pathways, and thus were less informative than the target analysis by ProTINA.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AVAILABILITY

MATLAB and R versions of ProTINA can be downloaded from Github repository (https://github.com/CABSEL/ProTINA).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We would like to thank Ziyi Hua for her assistance in preparing the R codes of PROTINA.

FUNDING

This work was supported by ETH Research Grant.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

REFERENCES

1. Imming,P., Sinning,C. and Meyer,A. (2006) Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.*, **5**, 821–834.
2. Schenone,M., Dančík,V., Wagner,B.K. and Clemons,P.A. (2013) Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.*, **9**, 232–240.
3. Isik,Z., Baldow,C., Cannistraci,C.V. and Schroeder,M. (2015) Drug target prioritization by perturbed gene expression and network information. *Sci. Rep.*, **5**, 17417.
4. Chua,H.N. and Roth,F.P. (2011) Discovering the targets of drugs via computational systems biology. *J. Biol. Chem.*, **286**, 23653–23658.
5. Lamb,J. (2007) The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer*, **7**, 54–60.
6. Iorio,F., Rittman,T., Ge,H., Menden,M. and Saez-Rodriguez,J. (2013) Transcriptional data: A new gateway to drug repositioning? *Drug Discov. Today*, **18**, 350–357.
7. Gardner,T.S. (2003) Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. *Science (80-.).*, **301**, 102–105.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

8. di Bernardo,D., Thompson,M.J., Gardner,T.S., Chobot,S.E., Eastwood,E.L., Wojtovich,A.P., Elliott,S.J., Schaus,S.E. and Collins,J.J. (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.*, **23**, 377–383.

9. Noh,H. and Gunawan,R. (2016) Inferring gene targets of drugs and chemical compounds from gene expression profiles. *Bioinformatics*, **32**, 2120–2127.

10. Chindelevitch,L., Ziemek,D., Enayetallah,A., Randhawa,R., Sidders,B., Brockel,C. and Huang,E.S. (2012) Causal reasoning on biological networks: Interpreting transcriptional changes. *Bioinformatics*, **28**, 1114–1121.

11. Martin,F., Thomson,T.M., Sewer,A., Drubin,D. a, Mathis,C., Weisensee,D., Pratt,D., Hoeng,J. and Peitsch,M.C. (2012) Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC Syst. Biol.*, **6**, 54.

12. Lefebvre,C., Rajbhandari,P., Alvarez,M.J., Bandaru,P., Lim,W.K., Sato,M., Wang,K., Sumazin,P., Kustagi,M., Bisikirska,B.C., *et al.* (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.*, **6**, 1–10.

13. Laenen,G., Thorrez,L., Börnigen,D. and Moreau,Y. (2013) Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Mol. Biosyst.*, **9**, 1676.

14. Emig,D., Ivliev,A., Pustovalova,O., Lancashire,L., Bureeva,S., Nikolsky,Y. and Bessarabova,M. (2013) Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. *PLoS One*, **8**.

15. Woo,J.H., Shimoni,Y., Yang,W.S., Subramaniam,P., Iyer,A., Nicoletti,P., Rodríguez Martínez,M., López,G., Mattioli,M., Realubit,R., *et al.* (2015) Elucidating Compound Mechanism of Action by Network Perturbation Analysis. *Cell*, **162**, 441–451.

16. Bansal,M., Yang,J., Karan,C., Menden,M.P., Costello,J.C., Tang,H., Xiao,G., Li,Y., Allen,J., Zhong,R., *et al.* (2014) A community computational challenge to predict the activity of pairs of compounds. *Nat. Biotechnol.*, **32**, 1213–1222.

17. Magkoufopoulou,C., Claessen,S.M.H., Tsamou,M., Jennen,D.G.J., Kleinjans,J.C.S. and Van delft,J.H.M. (2012) A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo. *Carcinogenesis*, **33**, 1421–1429.

18. Kubicek,S., Gilbert,J.C., Fomina-yadlin,D., Gitlin,A.D. and Yuan,Y. (2012) Chromatin-targeting small molecules cause class-speci fi c transcriptional changes in pancreatic endocrine cells. *Proc Natl Acad Sci U S A*, 10.1073/pnas.1201079109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1201079109.

19. McDermott,J.E., Shankaran,H., Eisfeld,A.J., Belisle,S.E., Neuman,G., Li,C., McWeeney,S., Sabourin,C., Kawaoka,Y., Katze,M.G., *et al.* (2011) Conserved host response to highly pathogenic avian influenza virus infection in human cell culture, mouse and macaque model systems. *BMC Syst. Biol.*, **5**, 190.

20. Li,C., Bankhead,A., Eisfeld,A.J., Hatta,Y., Jeng,S., Chang,J.H., Aicher,L.D., Proll,S., Ellis,A.L., Law,G.L., *et al.* (2011) Host Regulatory Network Response to Infection with Highly Pathogenic H5N1 Avian Influenza Virus. *J. Virol.*, **85**, 10955–10967.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. Mitchell,H.D., Eisfeld,A.J., Sims,A.C., McDermott,J.E., Matzke,M.M., Webb-Robertson,B.J.M., Tilton,S.C., Tchitchek,N., Josset,L., Li,C., *et al.* (2013) A Network Integration Approach to Predict Conserved Regulators Related to Pathogenicity of Influenza and SARS-CoV Respiratory Viruses. *PLoS One*, **8**.

22. Menachery,V.D., Eisfeld,A.J., Schäfer,A., Josset,L., Sims,A.C., Proll,S., Fan,S., Li,C., Neumann,G., Tilton,S.C., *et al.* (2014) Pathogenic influenza viruses and coronaviruses utilize similar and contrasting approaches to control interferon-stimulated gene responses. *MBio*, **5**, 1–11.

23. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., *et al.* (2013) NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.*, **41**, 991–995.

24. Gentleman,R., Carey,V., Bates,D., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

25. Marbach,D., Lamparter,D., Quon,G., Kellis,M., Kutalik,Z. and Bergmann,S. (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods*, **13**, 366–370.

26. Cahan,P., Li,H., Morris,S.A., Lummertz Da Rocha,E., Daley,G.Q. and Collins,J.J. (2014) CellNet: Network biology applied to stem cell engineering. *Cell*, **158**, 903–915.

27. Kuleshov,M. V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A., *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.

28. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P., *et al.* (2015) STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

29. Noh,H., Ziyi,H. and Gunawan,R. (2016) Inferring Causal Gene Targets from Time Course Expression Data. *IFAC-PapersOnLine*, **49**, 350–356.

30. Lynch, D. R., (2005) Finite Difference Calculus, Numerical Partial Differential Equations for Enviornmental Scientists and Engineers: A first Practical Course, pp. 11–19

31. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.*, **33**, 1–20.

32. Whitlock,M.C. (2005) Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.*, **18**, 1368–1373.

33. Stolovitzky,G., Prill,R.J. and Califano,A. (2009) Lessons from the DREAM2 challenges: A community effort to assess biological network inference. *Ann. N. Y. Acad. Sci.*, **1158**, 159–195.

34. Prill,R.J., Marbach,D., Saez-Rodriguez,J., Sorger,P.K., Alexopoulos,L.G., Xue,X., Clarke,N.D., Altan-Bonnet,G. and Stolovitzky,G. (2010) Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PLoS One*, **5**.

35. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

36. Luo,W., Friedman,M.S., Shedden,K., Hankenson,K.D. and Woolf,P.J. (2009) GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, **10**, 161.

37. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., MacIejewski,A., Arndt,D., Wilson,M., Neveu,V., *et al.* (2014) DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, 1091–1097.

38. Zhu,F., Shi,Z., Qin,C., Tao,L., Liu,X., Xu,F., Zhang,L., Song,Y., Liu,X., Zhang,J., *et al.* (2012) Therapeutic target database update 2012: A resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.*, **40**, 1128–1136.

39. Günther,S., Kuhn,M., Dunkel,M., Campillos,M., Senger,C., Petsalaki,E., Ahmed,J., Urdiales,E.G., Gewiess,A., Jensen,L.J., *et al.* (2008) SuperTarget and Matador: Resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, 919–922.

40. Davis,A.P., Grondin,C.J., Johnson,R.J., Sciaky,D., King,B.L., McMorran,R., Wiegers,J., Wiegers,T.C. and Mattingly,C.J. (2017) The Comparative Toxicogenomics Database: Update 2017. *Nucleic Acids Res.*, **45**, D972–D978.

41. Szklarczyk,D., Santos,A., Von Mering,C., Jensen,L.J., Bork,P. and Kuhn,M. (2016) STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.*, **44**, D380–D384.

42. Watanabe,T., Kawakami,E., Shoemaker,J.E., Lopes,T.J.S., Matsuoka,Y., Tomita,Y., Kozuka-Hata,H., Gorai,T., Kuwahara,T., Takeda,E., *et al.* (2014) Influenza virus-host interactome screen as a platform for antiviral drug development. *Cell Host Microbe*, **16**, 795–805.

43. Cazzalini,O., Scovassi,A.I., Savio,M., Stivala,L.A. and Prosperi,E. (2010) Multiple roles of the cell cycle inhibitor p21CDKN1A in the DNA damage response. *Mutat. Res. - Rev. Mutat. Res.*, **704**, 12–20.

44. Zhan,Q. (2005) Gadd45a, a p53- and BRCA1-regulated stress protein, in cellular response to DNA damage. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, **569**, 133–143.

45. Kelman,Z. (1997) PCNA: structure, functions and interactions. *Oncogene*, **14**, 629–640.

46. Shao,S., Wang,Y., Jin,S., Song,Y., Wang,X., Fan,W., Zhao,Z., Fu,M., Tong,T., Dong,L., *et al.* (2006) Gadd45a Interacts with Aurora-A and Inhibits Its Kinase Activity *. **281**, 28943–28950.

47. Lindqvist,A., Lim,D., Lampson,M.A., Klompmaker,R., Freire,R., Clouin,C., Taylor,S.S. and Yaffe,M.B. (2008) Polo-like kinase-1 is activated by aurora A to promote checkpoint recovery. **455**, 119–124.

48. Toyoshima-morimoto,F., Taniguchi,E. and Shinya,N. (2001) Labelling of purified sGC cyclin B1 and targets it to the nucleus during prophase. **410**, 215–220.

49. Pearl,L.H., Schierz,A.C., Ward,S.E., Al-Lazikani,B. and Pearl,F.M.G. (2015) Therapeutic opportunities within the DNA damage response. *Nat. Rev. Cancer*, **15**, 166–180.

50. Deans,A.J. and West,S.C. (2013) DNA interstrand crosslink repair and cancer. *Nat rev Cancer*, **11**, 467–480.

51. Andreassen,P. and Ren,K. (2009) Fanconi Anaemia Proteins, DNA Interstrand Crosslink Repair Pathways, and Cancer Therapy. *Curr. Cancer Crug Targets*, **9**, 101–117.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

52. de Chassey,B., Meyniel-Schicklin,L., Aublin-Gex,A., Navratil,V., Chantier,T., André,P. and Lotteau,V. (2013) Structure homology and interaction redundancy for discovering virus–host protein interactions. *EMBO Rep.*, **14**, 938–944.

53. Shaw,M.L., Stone,K.L., Colangelo,C.M., Gulcicek,E.E. and Palese,P. (2008) Cellular proteins in influenza virus particles. *PLoS Pathog.*, **4**, 1–13.

54. Matsuoka,Y., Matsumae,H., Katoh,M., Eisfeld,A.J., Neumann,G., Hase,T., Ghosh,S., Shoemaker,J.E., Lopes,T.J., Watanabe,T., *et al.* (2013) A comprehensive map of the influenza A virus replication cycle. *BMC Syst. Biol.*, **7**, 97.

55. Braakman,I., Hoover-Litty,H., Wagner,K.R. and Helenius,A. (1991) Folding of influenza hemagglutinin in the endoplasmic reticulum. *J. Cell Biol.*, **114**, 401–411.

56. Rodriguez,A., Perez-Gonzalez,A. and Nieto,A. (2007) Influenza Virus Infection Causes Specific Degradation of the Largest Subunit of Cellular RNA Polymerase II. *J. Virol.*, **81**, 5315–5324.

57. Rudnicka,A. and Yamauchi,Y. (2016) Ubiquitin in influenza virus entry and innate immunity. *Viruses*, **8**, 1–15.

58. Varga,Z.T., Grant,A., Manicassamy,B. and Palese,P. (2012) Influenza Virus Protein PB1-F2 Inhibits the Induction of Type I Interferon by Binding to MAVS and Decreasing Mitochondrial Membrane Potential. *J. Virol.*, **86**, 8359–8366.

59. Hale,B.G., Randall,R.E., Ortin,J. and Jackson,D. (2008) The multifunctional NS1 protein of influenza A viruses. *J. Gen. Virol.*, **89**, 2359–2376.

60. Shoemaker,J.E., Fukuyama,S., Eisfeld,A.J., Muramoto,Y., Watanabe,S., Watanabe,T., Matsuoka,Y., Kitano,H. and Kawaoka,Y. (2012) Integrated network analysis reveals a novel role for the cell cycle in 2009 pandemic influenza virus-induced inflammation in macaque lungs. *BMC Syst. Biol.*, **6**.

61. Cho,H., Ahn,S.H., Kim,K.M. and Kim,Y.K. (2013) Non-structural protein 1 of influenza viruses inhibits rapid mRNA degradation mediated by double-stranded RNA-binding protein, staufen1. *FEBS Lett.*, **587**, 2118–2124.

62. Li,N., Parrish,M., Chan,T.K., Yin,L., Rai,P., Yoshiyuki,Y., Abolhassani,N., Tan,K.B., Kiraly,O., Chow,V.T.K., *et al.* (2015) Influenza infection induces host DNA damage and dynamic DNA damage responses during tissue regeneration. *Cell. Mol. Life Sci.*, **72**, 2973–2988.

63. Elbahesh,H., Cline,T., Baranovich,T., Govorkova,E.A., Schultz-Cherry,S., and Russell,C.J. (2014) Novel Roles of Focal Adhesion Kinase in Cytoplasmic Entry and Replication of Influenza A Viruses. J. Virol. 88, 6714–6728.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE AND FIGURES LEGENDS

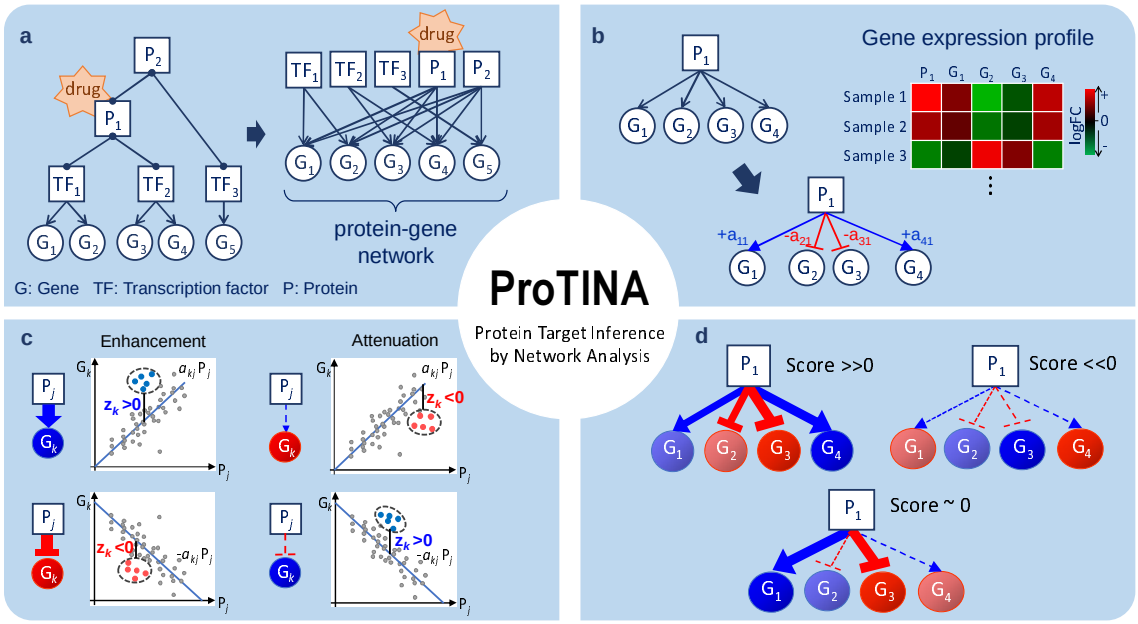

Figure 1. Protein target prediction by ProTINA. (a) The protein-gene network describes direct and indirect regulations of gene expression by transcription factors (TF) and their protein partners (P), respectively. A drug interaction with a protein is expected to cause differential expression of the downstream genes in the PGN. (b) Based on a kinetic model of gene transcriptional process, PROTINA infers the weights of the protein-gene regulatory edges, denoted by a_{kj}, using gene expression data. The variable a_{kj} describes the regulation of protein j on gene k, where the magnitude and sign of a_{kj} indicate the strength and mode (+a_{kj}: activation, -a_{kj}: repression) of the regulatory interaction, respectively. (c) A candidate protein target is scored based on the deviations in the expression of downstream genes from the PGN model prediction (P_j: log2FC expression of protein j, G_k: log2FC expression of gene k). The colored dots in the plots illustrate the log2FC data of a particular drug treatment, while the lines show the predicted expression of gene k by the (linear) PGN model. The variable z_k denotes the z-score of the deviation of the expression of gene k from the PGN model prediction. A drug-induced enhancement of protein-gene regulatory interactions is indicated by a positive (negative) z_k in the expression of genes that are activated (repressed) by the protein (i.e. $a_{kj}z_k > 0$). Vice versa, a drug-induced attenuation is indicated by a negative (positive) z_k in the expression of genes that are activated (repressed) by the protein (i.e. $a_{kj}z_k < 0$). (d) The score of a candidate protein target is determined by combining the z-scores of the set of regulatory edges associated with the protein in the PGN. A positive (negative) score indicates a drug-induced enhancement (attenuation). The larger the magnitude of the score, the more consistent is the drug induced perturbations (enhancement/attenuation) on the protein-gene regulatory edges.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

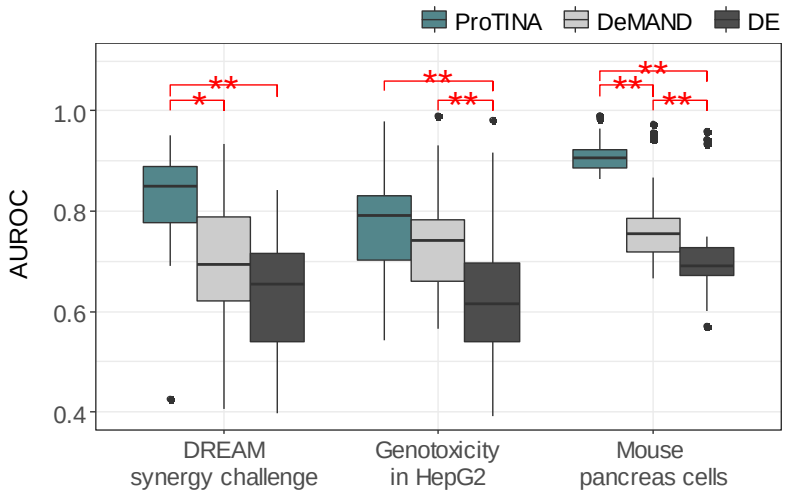

Figure 2. Prediction of known targets of drugs. AUROCs of protein target predictions from ProTINA, DeMAND and DE methods for the NCI-DREAM drug synergy (human B-cell lymphoma), the compound genotoxicity (human HepG2) and the chromatin targeting study (mouse pancreatic cell) datasets (*: p-value < 0.01, **: p-value <0.001 by paired t-test).

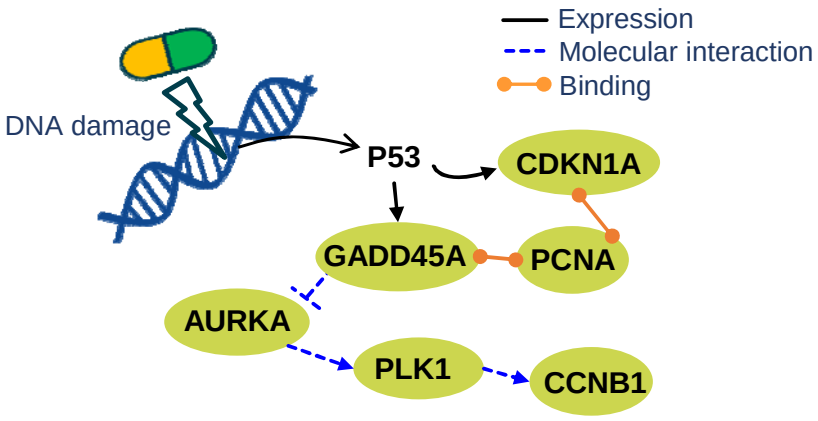

Figure 3. Canonical p53 DNA damage response pathway. In response to DNA damage, GADD45A, CDKN1A, PCNA are activated, while AURKA, CCNB1, and PLK1 proteins are inhibited (15).

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

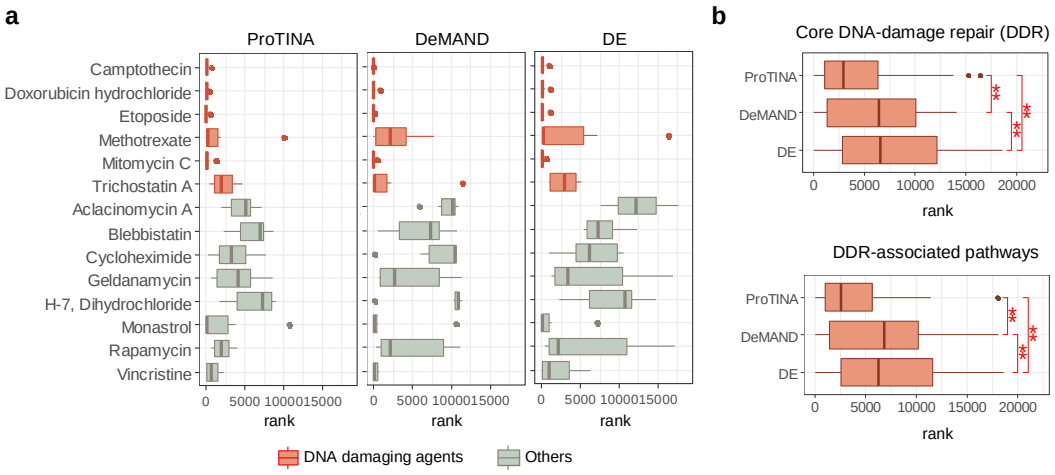

Figure 4. Mechanism of action of compounds based on target predictions by ProTINA. (a) The rank distribution of the canonical p53 DNA damage response proteins in the drug target predictions of PROTINA, DeMAND and DE for the NCI-DREAM drug synergy dataset. (B) The rank distribution of proteins involved in the core DNA-damage repair (DDR) and DDR-associated pathways (49) in the target predictions of PROTINA, DeMAND, and DE for the DNA damaging compounds in the NCI-DREAM drug synergy study (**: p-value <0.001 by Wilcoxon signed rank tests).

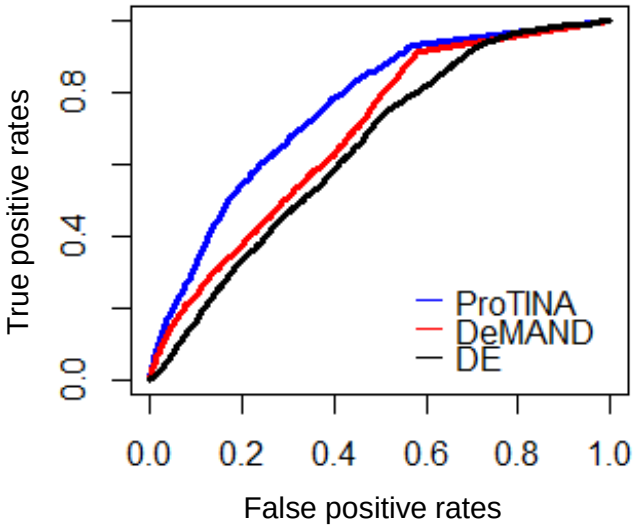

Figure 5. Prediction of targets of influenza A virus. The receiver operative characteristic curves give the true positive rate versus the false positive rate relationship of the protein target predictions from ProTINA, DeMAND, and DE against proteins that co-immunoprecipitate with influenza A viral proteins. The AUROCs for ProTINA, DeMAND and DE analysis are 0.758, 0.687 and 0.647, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/175364; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

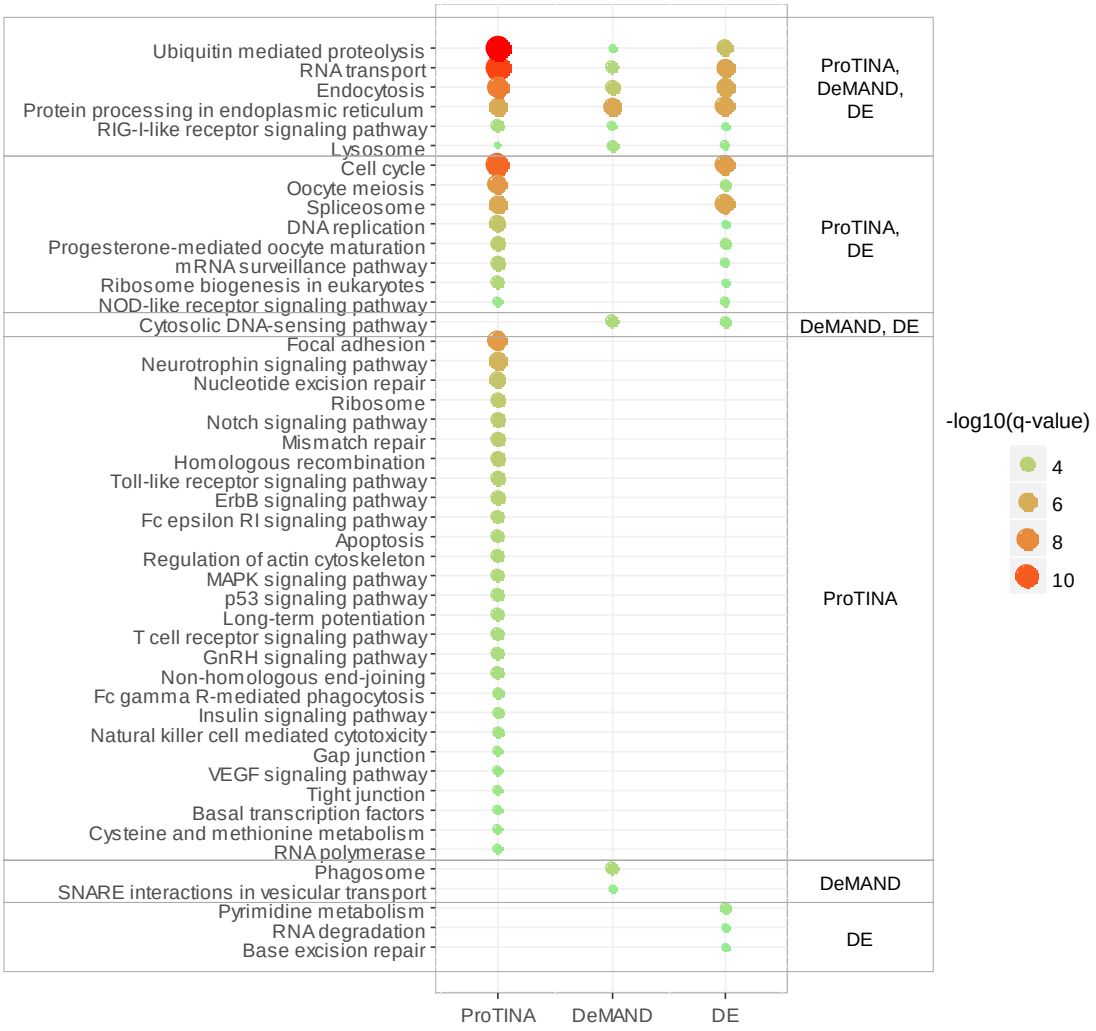

Figure 6. Gene set enrichment analysis for KEGG pathways for the influenza A protein target predictions from ProTINA, DeMAND, and DE. The size and color of dots correspond to –log 10 scale of the *q*-values. Only pathways with *q*-value < 0.01 are shown.