SALMON: inferring protein targets of compounds by network-based analysis of gene transcriptional profiles

Heeju Noh[*,†], Jason E. Shoemaker[‡,§] and Rudiyanto Gunawan[*,†]

We present SALMON, a method for identifying protein targets of compounds using cell-type specific protein-gene networks and gene transcriptional profiles. For benchmark datasets from three drug treatment studies, SALMON was able to provide highly accurate target predictions, with an average area under receiver operating characteristic of 0.82. SALMON was also able to reveal the mechanism of action of DNA-damaging compounds in NCI-DREAM drug synergy study with high sensitivity and specificity.

[*] Institute for Chemical and Bioengineering, ETH Zurich, Zurich, Switzerland, [†]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [‡]Department of Chemical and Petroleum Engineering, University of Pittsburgh, Pennsylvania, USA, and [§]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pennsylvania, USA. Correspondence should be addressed to R.G. (rudi.gunawan@chem.ethz.ch)

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 molecular structure whose interaction with the compound is connected to the compound's effects[1]. In this study, we treat proteins, particularly transcriptional factors (TFs) and their interacting protein partners, as the drug targets and transcriptional expressions as the drug 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. However, a drawback of using gene expression profiling for target discovery is that the data give only indirect evidence of the drug action. As illustrated in **Fig. 1a**, the interactions between a compound and its protein target(s) are expected to result in the differential expression of genes that are regulated by these proteins. The expression of the protein targets themselves may not – and often do not – change because of the drug's actions[3]. Consequently, the compound target prediction using gene expression profiles requires computational methods, taking into account the network of gene regulatory interactions, to delineate the (upstream) targets from the (downstream) effects.

Existing computational methods of gene expression analysis for compound target identification 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 likeness between the differential gene expression from a drug of interest and those from reference compounds or experiments with known targets, is used to indicate

similarity in the molecular targets and thereby the MoA. 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 gene regulatory network (GRN) and protein-protein interaction network (PIN). A number of network-based analytical methods relied on kinetic modeling of the GRN to infer the network perturbations caused by a drug treatment[7-9]. Several network-based analytical methods used statistical analysis to score potential drug targets based on the differential expression of genes in the network that are interacting with 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 prioritizations[3,13,14]. More recent methods combined different types of cellular networks. Notably, a method called Detecting Mechanism of Action by Network Dysregulation (DeMAND) combines GRN and PIN to create a molecular interaction network, where the drug targets are scored based on drug-induced alterations in the joint gene expression distribution between two connected nodes 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 leveraged comprehensive maps of protein-protein and protein-DNA interactions to construct, when possible, a tissue or cell type-specific protein-gene network

(PGN). Our method, called Systems Analysis and Learning for inferring Modifiers of Networks (SALMON), considers a PGN with directed edges (see **Fig. 1a**), describing direct and indirect gene transcriptional regulation by TFs and their protein partners. The edge weights in the PGN are determined by applying regularized regression (ridge regression) using the gene expression data, based on a kinetic model of the gene transcriptional process (see **Fig. 1b** and **online method**). Here, a positive weight indicates gene activation, while a negative weight implies gene repression. Because of the underlying kinetic model, SALMON 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 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 PGN 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 drug sample, a target is scored based on the overall enhancement and/or attenuation of its regulatory influence on the downstream genes (see **Fig. 1d** and **online method**). Thus, a protein 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.

We tested SALMON's performance in predicting drug targets using gene expression data from three drug treatment studies using 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 SALMON to the state-of-the-art network-based analytical method DeMAND[15], and to traditional differential expression (DE) analysis (see **online method**). For the analysis of the first two datasets, we constructed human cell-type specific PGNs by combining human PIN from STRING[19] and Enrichr database[20] and cell-type specific protein-DNA networks from Regulatory Circuit resource[21]. Meanwhile, for the construction of mouse pancreatic cell-specific PGN, we used PIN from STRING[19] and mouse protein-DNA interactions from CellNet[22] (see **online method**).

In assessing the performance of SALMON and the other methods, we compared the ranked list of protein target prediction for each compound with the reference drug targets compiled from the literature (see **online method** and **Supplementary material 1**). More specifically, we computed the area under the receiver operating characteristic curve (AUROC), i.e. the area under the true positive rate versus the false positive rate curve, where a higher AUROC value indicates a more accurate target prediction. **Fig. 2a** (also see **Supplementary Table S1-3**) summarizes the AUROCs of the target predictions from SALMON, DeMAND, and DE analysis, showing SALMON significantly outperforming DeMAND and DE analysis in all three studies. Here, the drug target predictions from DE analysis had the lowest AUROCs with an overall average below 0.67. Meanwhile, the target predictions of DeMAND were slightly better than the DE analysis, averaging at 0.74 for the three datasets. Meanwhile, SALMON gave the highest average AUROCs among the methods with an average of 0.82.

Besides high AUROCs, SALMON 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 SALMON's ability to reveal the compound MoA, we focused on the canonical p53 DNA damage response pathway[15], as illustrated in **Fig. 2b**. 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)[23,24]. In turn, CDKN1A and GADD45A – through their interactions with Proliferating Cell Nuclear Antigen (PCNA) – regulate the DNA replication and repair process[25]. GADD45A also inhibits the catalytic activity of Aurora Kinase A (AURKA)[26], leading to a lowered activation of Polo-like Kinase 1 (PLK1) and Cyclin B1 (CCNB1) in a phosphorylation cascade[27,28]. As shown in **Fig. 2c**, except for trichostatin A, the six proteins in the canonical p53 pathway above were ranked highly by SALMON among the genotoxic compounds (median rank <500) in the dataset, 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 SALMON predictions. Equally important, SALMON 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 SALMON 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. 2b**.

As illustrated in **Fig. 2c**, DeMAND and DE analysis also performed reasonably well in predicting the compounds' MoA. But, the directions of the drug perturbations were not predicted by DeMAND, and those from DE analysis were not always consistent with the expected response to DNA damage (see **Supplementary Table S5-6**). 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[29] (see **Supplementary material 2**). As depicted in **Fig. 2d**, SALMON ranked these proteins much higher than DeMAND and DE analysis, with DE performing the poorest among the methods considered. Moreover, in comparison to DeMAND and DE analysis, SALMON 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[30]. 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 Cancer Type 1 Susceptibility Protein (BRCA1) and RAD51 Recombinase (RAD51)[31]. SALMON analysis assigned FANCD2, FANCI, BRCA1, and RAD51 among the top 100 protein targets for mitomycin C, and not for the other DNA damaging agents (see **Supplementary Table S7**). The specific activation of the fanconi anemia pathway by mitomycin C was not detected by DeMAND or DE analysis. Thus, SALMON provided more sensitive and specific indications for the mechanism of action of compounds than DeMAND and DE.

In summary, SALMON is a novel and highly effective network-based analytical method for inferring the protein targets of compounds from gene expression profiling data. Using gene expression profiles of drug treatments, SALMON generates protein target scores, whose magnitudes reflect the confidence that the drug interacts with a particular protein and whose signs indicate how the drug alters the gene regulatory activity of its targets (enhancement or attenuation). The application of SALMON to gene expression profiles from three drug treatment studies demonstrated the capability of SALMON in predicting targets and MoA of compounds with high sensitivity and specificity across different cell types and species.

METHODS

Protein-Gene Network

The protein-gene network (PGN) is a bipartite graph with directed edges, pointing from a protein to a gene. The edges describe the regulation of gene expression by transcription factors (TFs) and their protein partners. As illustrated in **Fig. 1a**, the PGN is constructed by combining two types of networks, namely the TF-gene network and protein-protein interaction network (PIN). For the construction of human tissue-specific PGNs, we relied on the Regulatory Circuit resource[21] that provides 394 cell type and tissue-specific TF-gene interactions. In the analysis of the NCI-DREAM drug synergy dataset, we used the TF-gene network of human lymphoma cells, while for the genotoxic compound study dataset, we employed the TF-gene network of pleomorphic hepatocellular carcinoma cells. Here, we included only TF-gene interactions with a confidence score greater than 0.1. Meanwhile, for the mouse gene expression dataset, we obtained the mouse pancreatic TF-gene interactions from CellNet[22]. In the construction of human and mouse PGNs, any TF-gene interactions involving unmeasured genes were excluded. In summary, the TF-gene network for human lymphoma and hepatocellular carcinoma cell lines included 31,392 and 3,868 interactions (edges) among 515 TFs – 5,153 genes and 413 TFs – 953 genes, respectively. On the other hand, the mouse pancreatic PGN contained 2,922 interactions, involving 95 TFs and 588 genes.

For the PIN in human, we combined the information from two databases, namely Enrichr[20] and STRING[19]. Meanwhile, for mouse pancreatic cells, we used STRING database. We used the PINs to identify protein partners of the TFs, defined as proteins that are within a network distance of 2 from the TFs in the PIN. When using STRING, we included all direct protein partners of TFs, and proteins with a network distance of 2 from TFs with a score larger

than 500. For human lymphoma cells, we found 10,649 protein partners for a subset of 499 TFs (out of 515 TFs), while for human hepatocytes, we found 10,488 protein partners for a subset of 403 TFs (out of 413 TFs). For mouse pancreatic cells, we identified 6,598 protein partners for a subset of 89 TFs (out of 95 TFs).

Finally, in the construction of the PGNs, 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-specific PGN for human lymphoma cells included 21,490,181 regulatory edges among 11,161 TFs/proteins and 5,153 genes. For hepatocellular carcinoma cells, the cell-specific PGN comprised 3,726,671 edges among 10,893 TFs/proteins and 953 genes. For mouse pancreatic cells, the cell-specific PGN consisted of 1,418,067 edges among 6,661 TFs/proteins and 588 genes.

Gene Transcription Model

The edges in the PGN 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[9,32]. The inference of the edge weights is based on an ordinary differential equation (ODE) model of the mRNA production of a gene:

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

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. The model is based on the assumption that the regulatory activity of a protein is concomitant with its mRNA expression.

While the regulatory edges in the model above usually describe TF-gene interactions, in SALMON, 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:

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

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 *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 *j* (a non-zero r_j). The model in Equation (2) can be simplified into:

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

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), and thus the inference of the weights could be carried out using the same procedure as in DeltaNet[9,32].

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:

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

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 dysregulation of the expression of gene k. As detailed below, SALMON relies on the magnitude and directions of such network dysregulations to identify proteins with altered gene regulatory activity.

As previously discussed in ref. 32, dynamical information of the PGN contained within time-series gene expression profiles, could greatly improve the inference of the edge weights. 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[33]. In summary, the estimation of edge weights in SALMON 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-specific PGN, $\mathbf{C}_{R_k}$ 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 PGN pointing to gene k, and $\mathbf{p}_k$ is the $1\times m$ vector of dysregulation impacts of gene k over m samples.

In SALMON, 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 $\mathbf{y} = [\mathbf{C}_k \;\; \mathbf{S}_k]^T$, $\mathbf{X} = [\,[\mathbf{C}_{R_k} \;\; \mathbf{S}_{R_k}]^T, \;\; [I_m \;\; \mathbf{0}]^T\,]$, $\beta = [\mathbf{A}_k \;\; \mathbf{P}_k]^T$. Before applying the ridge regression, we normalized the vectors of log2FCs and slopes to have a unit norm. In the applications of SALMON, 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[34] for both the MATLAB and R versions of SALMON.

Drug target scoring

In SALMON, 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.

$$\mathbf{r}_k = \mathbf{C}_k - \mathbf{A}_k\mathbf{C}_{R_k} \qquad (9)$$

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

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

where $\bar{r}_{lk}$ 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 PGN, as follows: (ref. 35)

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

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 PGN model. In contrast, a negative $w_{kj}z_{ki}$ indicates an attenuation in 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 PGN 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}.

Implementation of DeMAND and differential expression analysis

For DeMAND analysis, we employed the R subroutines from http://califano.c2b2.columbia.edu/demand. Following the procedure of DeMAND[15], we computed the RMA (Robust Multi-array Average) normalized gene expression values as inputs to the analysis. In addition, we used the same cell-specific PGNs in DeMAND as in SALMON. 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 PGN. 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 PGN, as described in section **Gene expression data** below. Here, we used the 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) following the procedure adopted in DREAM

challenges[36,37], i.e. the area under the plot of true positive rate against false positive rate. For each method and each drug treatment, we generated a ranked list of protein targets according to decreasing magnitudes of the protein scores in SALMON, increasing *p*-values of network dysregulation in DeMAND, and increasing magnitudes of log2FC gene expression in DE analysis.

Gene expression data

For NCI-DREAM drug synergy data, we obtained the raw *Affymetrix Human Genome U219* microarray data from Gene Expression Omnibus (GEO) database[38] (accession number: GSE51068). The raw data were first normalized and transformed into log2-scaled expressions using *justRMA* function in the *affy* package of Bioconductor[39]. Then, the log2FC differential expressions and their statistical significance were calculated using a linear fit model and empirical Bayes method in the *limma* package of Bionconductor. Three samples from the drug treatment using low concentration of Aclacinomycin were dropped because the log2FC expressions were close to 1 and 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 *p*-value over the samples.

The raw microarray data from genotoxicity study[17] in human HepG2 cell line were obtained from GEO database (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 *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 averaged *p*-value across all samples was chosen.

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

Compilation of known drug targets

The reference protein targets of the drugs were compiled from 5 different public databases of chemical-protein interactions: DrugBank[40], Therapeutic Target Database (TTD)[41], MATADOR[42], Comparative Toxicogenomics Database (CTD)[43], and STITCH[44]. DrugBank and TTD provide information on the mechanism of drug actions as well as the proteins that have physical binding interactions with drugs. Meanwhile, MATADOR, CTD, and STITCH give comprehensive interactions between proteins and chemical compounds, curated from text mining and experimental evidences. When retrieving the protein targets of drugs from these databases, we only collected proteins that directly bind to the queried drugs. The reference drug targets for each dataset in this study are provided in **Supplementary material 1**.

Code availability

MATLAB and R versions of SALMON can be downloaded from the following website: https://github.com/CABSEL/SALMON.

SUPPLEMENTARY INFORMATION

supplementary_tableS1-7.pdf, supplementary_material1.xlsx, and supplementary_material2.xlsx

ACKNOWLEDGEMENTS

We would like to thank Ziyi Hua for her assistance in preparing the R codes of SALMON. This work was supported by ETH Research Grant.

AUTHOR CONTRIBUTIONS

H.N. developed the algorithms with inputs from the other authors. H.N. performed data collection, processing and analysis. R.G. and J.S. provided guidance throughout the study. H.N. generated and prepared the results including figures and supplementary materials. H.N. prepared MATLAB and R packages for the work. H.N, J.S. and R.G. wrote the manuscript.

COMPLETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Imming, P., Sinning C. & Meyer A. *Nat. Rev. Drug Discov.* **5**, 821–34 (2006).

2. Schenone, M., Dančík V., Wagner B. K. & Clemons P. *Nat. Chem. Biol.* **9**, 232–40 (2013).

3. Isik, Z., Baldow C., Cannistraci C. V. & Schroeder M. 2015. *Sci. Rep.* https://doi.org/10.1038/srep17417 (2015).

4. Chua, H. N. & Roth F. P. *J. Biol. Chem.* **286**, 23653–23658 (2011)

5. Lamb, J. *Nature* **7**, 54–60 (2007).

6. Iorio, F., Rittman T., Menden H. Ge, M. & Saez-Rodriguez J. *Drug Discov. Today* **18**, 350–357 (2013).

7. Gardner, T. S. *Science* **301**, 102–105 (2003).

8. Di Bernardo D. *et al*. *Nat. Biotechnol*, **23**, 377–383 (2005).

9. Noh, H. & Gunawan R. *Bioinformatics* **32**, 2120-2127 (2016)

10. Chindelevitch, L. *et al*. *Bioinformatics* **28**, 1114–1121 (2012).

11. Martin, F. *et al*. *BMC Syst. Biol.* https://doi.org/10.1186/1752-0509-6-54 (2012).

12. Lefebvre, C. *et al*. *Mol. Syst. Biol.* https://doi.org/10.1038/msb.2010.31 (2010).

13. Laenen, G., Thorrez L., Börnigen D. & Moreau Y. *Mol. BioSyst.* **9**, 1676-1685 (2013).

14. Emig, D. *et al*. *PLoS ONE* https://doi.org/10.1371/journal.pone.0060618 (2013).

15. Woo, J. H. *et al*. *Cell* **162**, 441–451 (2015).

16. Bansal, M. *et al*. *Nat. Biotechnol.* **32**, 1213-1222 (2014).

17. Magkoufopoulou, C. *et al*. *Carcinogenesis* **33**, 1421–1429 (2012).

18. Kubicek, S. *et al*. *P. Natl. Acad. Sci. USA* **109**, 5364–5369 (2012).

19. Szklarczyk, D. *et al*. *Nucleic Acids Res.* **43**, D447–D452 (2015).

20. Kuleshov M. V. *et al*. *Nucleic Acids Res*. **8**, W90-W907 (2016).

21. Marback, D. *et al*. *Nat. Methods* **13**, 366-370 (2016)

22. Cahan, P. *et al*. *Cell* **158**, 903–915 (2014).

23. Cazzalini, O., Scovassi A. I., Savi M. Stivala L. A. & Prosperi E. *Mutat. Res.* **704**, 12–20 (2010).

24. Zhan, Q. *Mutat. Res.* **569**, 133–143 (2005).

25. Kelman, Z. *Oncogene* **14**, 629–640 (1997).

26. Shao, S. *et al*. *J. Biol. Chem.* **281**, 28943–28950 (2006).

27. Macůrek, L. *et al*. *Nature* **455**, 119–123 (2008).

28. Toyoshima-Morimoto, F., Taniguchi E., Shinya N., Iwamatsu A., and Nishida E. *Nature* **410**, 215–220 (2001).

29. Pearl, L. H., Schierz A. C., Ward S. E., Al-lazikani B. & Pearl F. M. G. *Nat. Rev. Cancer* **15**, 166–180 (2015).

30. Deans, A. J. & West S. C. *Nat. Rev. Cancer* **11**, 467–480 (2011).

31. Andreassen, P. & Ren K. Curr. Cancer Drug Tar. **9**, 101–117 (2009).

32. Noh, H., Ziyi H. & Gunawan R. *IFAC PapersOnLine* **49**, 350–356 (2016).

33. Lynch, D. R., Springer, *USA* (2005).

34. Friedman, J,, Hastie T. & Tibshirani R.. *J. Stat. Softw.* **33**, 1-22 (2010).

35. Whitlock, M. C. *J. Evolution. Biol.* **18**, 1368–1373 (2005).

36. Stolovitzky, G., Prill R. J. & Califano A. *Ann. NY Acad. Sci.* **1158**, 159–195 (2009).

37. Prill, R. J. *et al*. *PLoS ONE* https://doi.org/10.1371/journal.pone.0009202 (2010).

38. Barrett, T. *et al*. *Nucleic Acids Res.* **41**, 991–995 (2013).

39. Gentleman, R. *et al*. *Genome Biol.* https://doi.org/10.1186/gb-2004-5-10-r80 (2004).

40. Law, V. *et al*. *Nucleic Acids Res.* **42**, 1091–1097 (2014).

41. Zhu, F. *et al*. *Nucleic Acids Res*. **40**, 1128–1136 (2012).

42. Günther, S. *et al*. *Nucleic Acids Res.* **36**, 919–922 (2008).

43. Davis, A. P. *et al*. *Nucleic Acids Res.* **45**, D972-D978 (2017).

44. Szklarczyk, D. *et al*. *Nucleic Acids Res.* **44**, D380–D384 (2016).

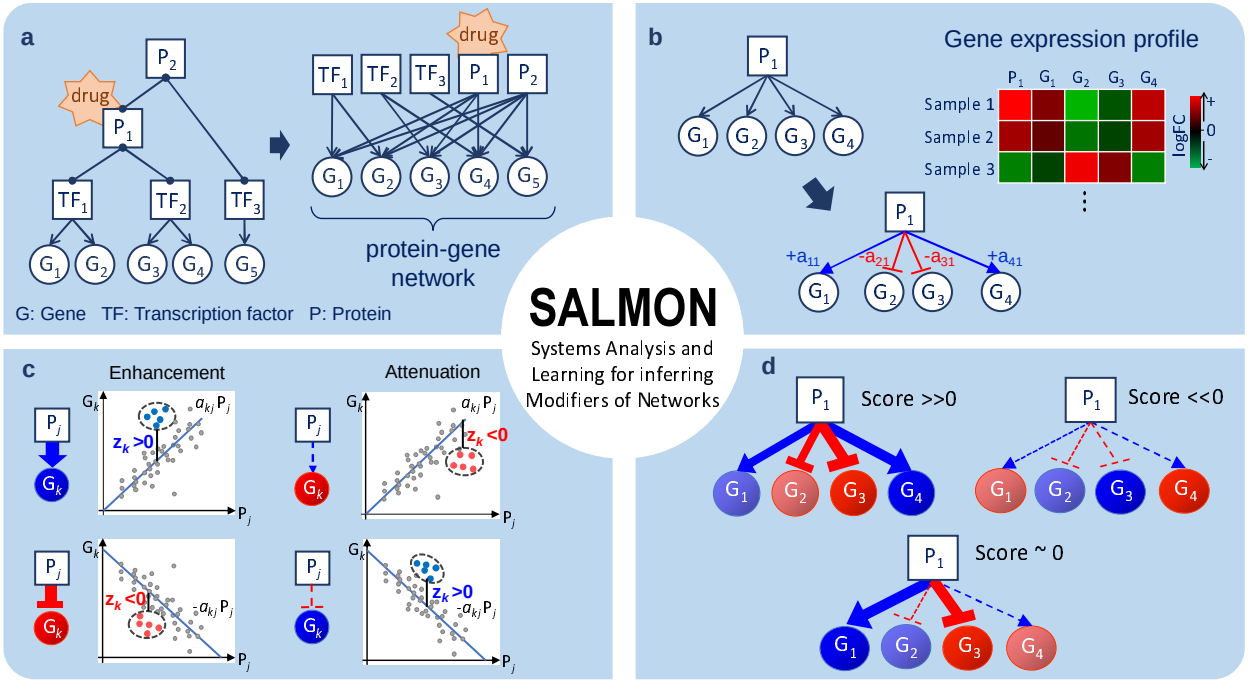

Figure 1. Protein target prediction by SALMON. (a) The protein-gene network describes direct and indirect regulations of gene expression by transcription factors (TF) and their protein partners (P), respectively. A drug interaction with a protein is expected to cause differential expression of the downstream genes in the PGN. (b) Based on a kinetic model of gene transcriptional process, SALMON 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.

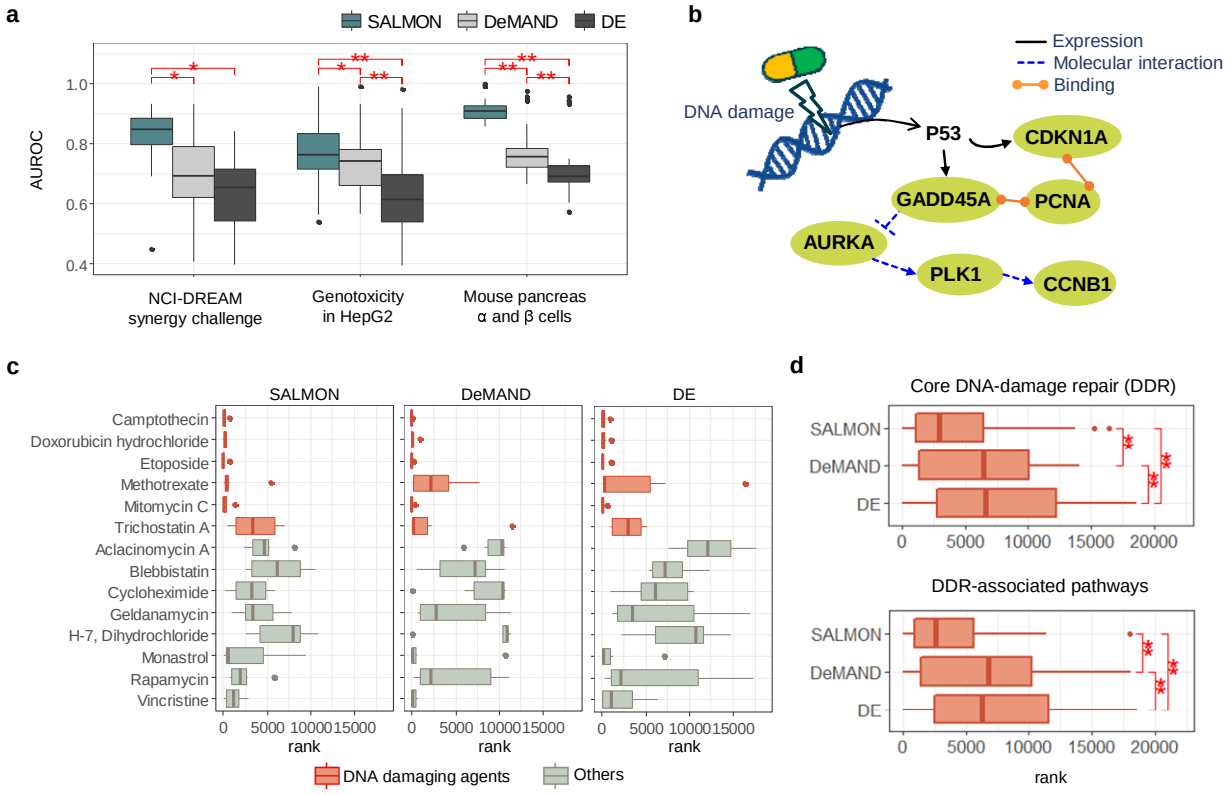

Figure 2. Protein target and MoA prediction by SALMON. (a) AUROCs of protein target predictions from SALMON, DeMAND and DE methods for the NCI-DREAM drug synergy (human B-cell lymphoma), the compound genotoxicity (human HepG2) and the chromatin targeting study (mouse pancreatic cell) datasets (*: p-value < 0.01, **: p-value <0.001 by paired t-test). (b) Canonical p53 DNA damage response pathway: GADD45A, CDKN1A, PCNA are activated, while AURKA, CCNB1, and PLK1 proteins are inhibited in response to DNA damage[15]. (c) The rank distribution of the canonical p53 DNA damage response proteins in the drug target predictions of SALMON, DeMAND and DE for the NCI-DREAM drug synergy dataset. (d) The rank distribution of proteins involved in the core DNA-damage repair (DDR) and DDR-associated pathways[29] in the target predictions of SALMON, DeMAND, and DE for the DNA damaging compounds in the NCI-DREAM drug synergy study (**: p-value <0.001 by Wilcoxon signed rank tests).