DRUG-INDUCED mRNA SIGNATURES ARE ENRICHED FOR THE MINORITY OF GENES THAT ARE HIGHLY HERITABLE

TIANXIANG GAO[1*], PETTER BRODIN[2,3*], MARK M DAVIS[3,4], VLADIMIR JOJIC[1*]

[1] *Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA*

[2] *Science for Life laboratory, Department of Medicine, Solna Karolinska Institutet SE-171 76 Stockholm, SWEDEN*

[3] *Department of Microbiology and Immunology, Stanford University School of Medicine,*

[4] *The Howard Hughes Medical Institute, Stanford University School of Medicine,*

∗ *E-mail: {tgao,vjojic}@cs.unc.edu, petter.brodin@ki.se*

The blood gene expression signatures are used as biomarkers for immunological and non-immunological diseases.[1] Therefore, it is important to understand the variation in blood gene expression patterns and the factors (heritable/non-heritable) that underlie this variation. In this paper, we study the relationship between drug effects on the one hand, and heritable and non-heritable factors influencing gene expression on the other. Understanding of this relationship can help select appropriate targets for drugs aimed at reverting disease phenotypes to healthy states. In order to estimate heritable and non-heritable effects on gene expression, we use Twin-ACE model on a gene expression dataset MuTHER,[2] measured in blood samples from monozygotic and dizygotic twins. In order to associate gene expression with drug effects, we use CMap[3,4] database. We show that, even though the expressions of most genes are driven by non-heritable factors, drugs are more likely to influence expression of genes, driven by heritable rather than non-heritable factors. We further study this finding in the context of a gene regulatory network. We investigate the relationship between the drug effects on gene expression and propagation of heritable and non-heritable factors through regulatory networks. We find that the decisive factor in determining whether a gene will be influenced by a drug is the flow of heritable effects supplied to the gene through regulatory network.

1. Introduction

In this paper, we examine a general question: whether a drug aiming to perturb a disease phenotype should target genes whose expression is dominated by heritable or non-heritable factors?

The expression level of a gene is determined by both heritable effects and non-heritable effects. We estimated those effects for expressions of 3245 genes from MuTHER twin database (in Figure 1) and found that the expression of most genes are driven by non-heritable effects.[5] At first, we would expect that a gene that is significantly impacted by non-heritable effects would be more likely to be affected by a drug – drugs could take advantage of such gene's environmentally driven variability.

Naturally, heritable factors also play an important role in drug response.[6–8] In our study, we find that the strong heritable effects on a gene's expression are predictive of whether drugs can influence this gene. A simple experiment in Section 3.2 uses CMap database to show that genes robust to non-heritable effects – hence, strongly driven by heritable effects – are more likely to be part of a drug influenced gene expression signature.

The result of this experiment led us to examine the first question in a broader context of gene regulatory network. Previous studies show that genes can pass the drug influence through
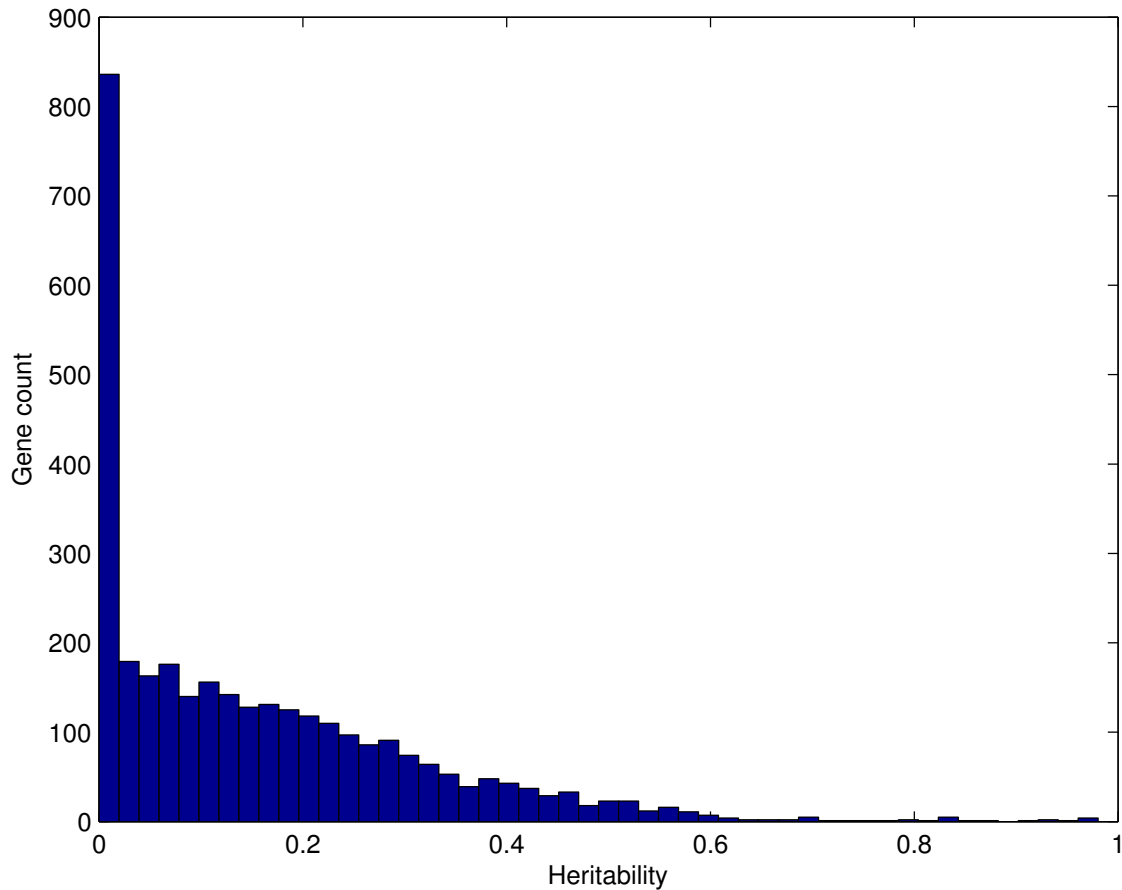
Fig. 1. The heritability estimation for 3245 genes from MuTHER twin database. Heritability is the ratio between variance driven by heritable effects and total variance.

the biological network to induce change in a target gene.[9,10] We deem such genes source genes, as they can be seen as sources of drug effect propagation in the network. Hence, we wondered whether we can predict the flow of drug influence through a regulatory network from source to target genes.

We say that there exists a **regulation flow** between a source gene and a target gene if there is a sequence of strong regulatory relationships leading from the source gene to the target gene. Naturally, highly variable source gene with a strong flow to target gene will induce variance in the target gene. We can estimate how much of the heritably or non-heritably driven variance is propagated from source to target. Consequently, we introduced quantitative measures of strength of heritable or non-heritable flow between source and target genes.

We call a regulation flow between a source and a target gene **drug influence flow** if there exists a drug influencing both genes. Hence we can pose a question: Does a strong heritable flow between source and a target imply existence of a drug influence flow between these two genes?

An experiment in Section 3.3 reveals the fact: strong heritable flows are predictive of drug influence flows. Equally importantly, strong non-heritable flows are *not* predictive of drug

influence flows. Hence, a drug influence flow leading to particular target gene is best identified through the regulation flows propagating substantial heritable effects.

Here, we provide a simple example of this relationship in Figure 2. In this example, both the source gene PI4KB and PIK3R5 have regulation flow to target gene INPP5F. However, heritable flow strength to INPP5F from PI4KB (red) is higher than heritable flow strength from PIK3R5, even though they have similar regulation flow strength. Our validation in CMap shows that there is no drug influencing both PIK3R5 and INPP5F; while the drug "phthalylsulfathiazole" is known to influence both PI4KB and INPP5F.
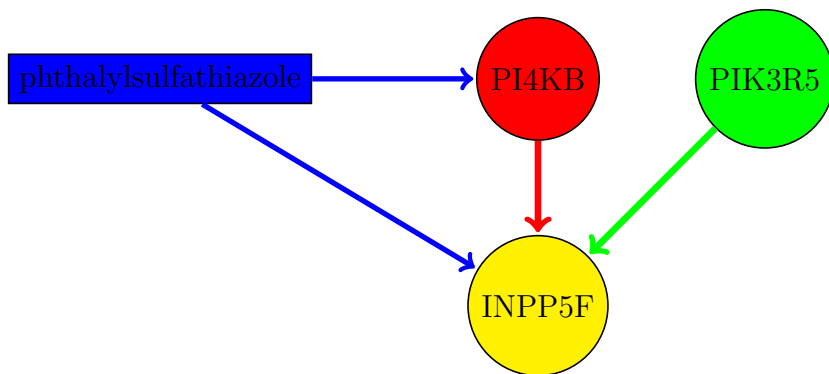

Fig. 2. An example of choosing the drug influence flow from our results. Yellow colored gene INPP5F is the target gene that we want to influence. The red regulation flow from PI4KB has regulation flow strength 0.3 and heritable flow strength of 0.1. The green regulation flow from PIK3R5 has regulation flow strength 0.3 and heritable flow strength of 0. The blue nodes is one of the drugs that is known to influence both PI4KB and INPP5F.

The rest of this paper is organized as follows: we introduce methods to find the drug influence flows in a specific network and Twin-ACE model for heritable effects estimation in Section 2. Supportive experiments and results are discussed in Section 3 and 4.

2. Methods

In this section, we discuss the methods that we use to recover the regulatory network and estimate heritable and non-heritable effects. In Section 2.1, we first introduce the directed acyclic graph gene regulatory network and the parameter estimation of a given network using linear model. In the second part, we introduce Twin-ACE model that we are using to estimate the impact of heritable and non-heritable effects on genes' expression, and the quantitative measures of the heritable and non-heritable flow strength in a regulatory network. Twin-ACE model in combination with regulatory network model can be seen as a mixed model of the joint data.[?]

2.1. *DAGRN and its estimation*

Directed Acyclic Gene Regulatory Network(DAGRN) Our method operates on a DAGRN, as this representation enables a straightforward way of calculating the regulatory

effects using linear regression of expression. Specifically, DAGRN is a graph of P nodes with no loops or undirected edges. Each node t in the graph is a random variable x_t represents the measurement of a gene expression value. It has the following conditional probability density:

$$x_t | \mu_t, \mathbf{W}, \sigma_t^2 \sim \mathcal{N}(\mu_t + \sum_{s \in pa(t)} w_{s,t} x_s, \sigma_t^2), \tag{1}$$

where $pa(t)$ is the set of parent nodes that link to node t. $\mathbf{W}$ is an adjacency matrix of size $P \times P$. Each entry $w_{s,t}$ indicates the strength of edge $s \to t$. μ_t is the local mean. We call σ_t^2 the residual variance of x_t.

Linear model estimation Given a gene expression matrix $\mathbf{X}$ of size $N \times P$, where N is the number of samples and P is the number of genes, we can estimate the parameters $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_P]$, $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_P]$ and $\mathbf{W}$.

Let us look at a specific gene. We use a vector $\mathbf{y}$ to denote the N observed samples for that gene. We denote regulators' gene expression matrix as $\mathbf{R}$. This is a matrix of size $N \times R$. Each of R columns is an observed expression levels corresponding to a parent node, a regulator. We can write the joint distribution for all the samples as:

$$p(\mathbf{y} | \mu, \mathbf{w}, \sigma) = \prod_{i=1}^{N} \mathcal{N}(\mu + \mathbf{r}_i \mathbf{w}, \sigma^2), \tag{2}$$

and $\mathbf{r}_k$ denotes k^{th} row of matrix $\mathbf{R}$. μ is the local mean for $\mathbf{y}$, σ^2 is the residual variance. We can derive the maximum likelihood estimation (MLE) update for parameters in (2) as :

$$\mu^{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{3}$$

$$\sigma^{\text{MLE}} = \frac{1}{N} ||\tilde{\mathbf{y}} - \mathbf{R}\mathbf{w}||_2^2 \tag{4}$$

$$\mathbf{w}^{\text{MLE}} = \underset{\mathbf{w}}{\text{argmin}} \, ||\tilde{\mathbf{y}} - \mathbf{R}\mathbf{w}||_2^2. \tag{5}$$

We write $\tilde{\mathbf{y}} = \mathbf{y} - \mu^{\text{MLE}} \mathbf{1}_N$ for convenience. The optimization problem (5) is a linear regression problem.

Stimulation-Response Matrix After acquiring the regulation weights for the network, we need to calculate the total regulation influence from any source node to any target node. Therefore, we introduce the Stimulation-Response matrix. To differentiate the notation from μ_t, which is the local mean of a gene expression x_t, we define $\alpha_t \triangleq \mathbb{E}[x_t]$ as the global mean of x_t. Suppose the drug treatment changes mean of the source gene x_s by $\Delta\alpha_s$, we want to know the response change $\Delta\alpha_t$ in the target gene. For convenience, we will call $\Delta\alpha_s$ **stimulation change** and $\Delta\alpha_t$ **response change**.

As our model is a Linear-Gaussian model, the stimulation change and response change are also linear:[11] $\Delta\alpha_t = S_{s,t}\Delta\alpha_s$. We call $\mathbf{S}$ Stimulation-Response matrix. It is a $P \times P$ matrix, where entry (s, t) indicates the response change $\Delta\alpha_t$ in the target gene t given unit stimulation change ($\Delta\alpha_s = 1$) in the source gene s. We can use Algorithm 1 to calculate the Stimulation-Response matrix. The computational complexity for this method is $O(P^2 V)$, V is the total edge number.

```
input   : W: adjacency matrix, P : total node number
output: S : stimulation-response matrix
initialize S ← 0_{P×P};
for  x ← 1 to P do
    S_{x,x} ← 1;
    for y ← nodes has directed edges link from x do
        for j ← 1 to P do
            S_{j,y} ← S_{j,y} + S_{j,x} * W_{x,y};
        end
    end
end
```

Algorithm 1: The algorithm for calculating Stimulation-Response matrix

We call the value in entry (s,t) of **S** as the **regulation flow strength** of the regulation flow $s \to t$. We only consider a regulation flows between source gene s and target gene t for which $S(s,t) > T$, where T is a certain threshold.

2.2. *Twin-ACE model and heritable/non-heritable effect estimation*

We are using the Twin-ACE model to estimate the heritable and non-heritable effects on the phenotypes. The standard ACE model for twin studies is based on the assumption that identical twins share their genes while fraternal twins share approximately half of their polymorphic gene sequences. ACE study design assumes presence of both monozygotic and dizygotic twins. In order to model the relatedness of the gene expression measurements in twins, we utilize the standard ACE model.[12] The main assumption underlying the ACE models is that the covariance of the gene expression in a twin pair can be decomposed into three contributions: 1) **A**dditive genetic component 2) **C**ommon environmental component, and 3) twin-specific **E**nvironmental component.

The additive genetic component is the only component that is dependent on whether the twins are identical (monozygotic, MZ) or fraternal (dizygotic, DZ). Identical twins share same genetic material and hence differences in their gene expressions are attributable to environmental factors. In contrast, fraternal twins, on average, share only half of their genetic sequences and hence differences in their gene expressions can be attributed to heritable or non-heritable factors. This observation motivates parametrization of phenotypic covariance in terms of additive components $\mathbf{A}_{\mathtt{MZ}} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $\mathbf{A}_{\mathtt{DZ}} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ reflecting the expectation of higher covariance among mono-zygotic twins to the extent the gene expression is heritable. In terms of notation, to indicate zygosity of a twin pair t_1 and t_2 we will use $\mathtt{zyg}(t_1, t_2)$, naturally $\mathtt{zyg}(t_1, t_2) \in \{\mathtt{MZ}, \mathtt{DZ}\}$.

In addition to additive genetic effects, we also model potential environment effects. We denote them as **C** and **E**. These effects are assumed to be independent of twins' genomes. Furthermore, the common environmental effects are assumed to be affecting both of the twins

in a family, and hence off-diagonal covariance terms are 1. The twin-specific environmental effects are assumed to have an independent effect on each of the twins, so the off-diagonal terms are 0. Hence, we have $\mathbf{C} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $\mathbf{E} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Estimation of ACE parameters In order to use the relatedness of the gene expression measurements in twins to estimate the heritable and non-heritable effects, we use Twin-ACE model. Suppose we have F families and each family has two twins. For a given twin pair (t_1, t_2) we will use $\mathbf{zyg}(t_1, t_2)$ to indicate whether the twin pair is monozygotic (MZ) or dizygotic (DZ). Similarly, we have the gene expression measurement $\mathbf{y}$ for $N = 2F$ twins. The joint distribution for all samples is:

$$p(\mathbf{y}|\mu, a, c, e) = \prod_{(t_1, t_2)} \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \boldsymbol{\Sigma}_{\mathbf{zyg}(t_1, t_2)}(a, c, e)\right), \tag{6}$$

where

$$\boldsymbol{\Sigma}_{\mathbf{zyg}(t_1, t_2)}(a, c, e) = a^2 \mathbf{A}_{\mathbf{zyg}(t_1, t_2)} + c^2 \mathbf{C} + e^2 \mathbf{E}.$$

We call a, c, e ACE parameters for gene expression vector $\mathbf{y}$. We will abbreviate covariances $\boldsymbol{\Sigma}_{\text{MZ}}, \boldsymbol{\Sigma}_{\text{DZ}}$ while acknowledging their dependence on parameters a, c, e.

In this model, μ is the population mean of the $\mathbf{y}$. To estimate the parameters a, c, e, we need to solve the follow optimization:

$$(a, c, e)^{\text{MLE}} = \underset{a,c,e}{\operatorname{argmax}} \left[-\frac{1}{2}(\mathbf{y} - \mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mu) - \frac{1}{2} \log |\boldsymbol{\Sigma}| \right]$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{zyg}(t_1, t_2)} & 0_{2\times 2} & 0_{2\times 2} & \dots & 0_{2\times 2} \\ 0_{2\times 2} & \boldsymbol{\Sigma}_{\mathbf{zyg}(t_3, t_4)} & 0_{2\times 2} & \dots & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} & \boldsymbol{\Sigma}_{\mathbf{zyg}(t_5, t_6)} & \dots & 0_{2\times 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{2\times 2} & 0_{2\times 2} & 0_{2\times 2} & \dots & \boldsymbol{\Sigma}_{\mathbf{zyg}(t_{2F-1}, t_{2F})} \end{bmatrix},$$

This optimization problem can be solved by Newton's method.[13] We define the heritable effects on a gene as a^2, and the non-heritable effects on a gene as $c^2 + e^2$.

Heritable/non-heritable flow strength The total variance propagated through a regulation flow can be decomposed into heritable flow and non-heritable flow. We define the **heritable flow strength** of a regulation flow $s \to t$ as: $G(s, t) = a_s^2 \times S(s, t)$. $G(s, t)$ indicates how much of heritably driven variance is propagated by the regulation flow from source gene s to target gene t. Similarly, we define **non-heritable flow strength** $E(s, t) = (c_s^2 + e_s^2) \times S(s, t)$ as the amount of non-heritably driven variance propagated by the regulation flow.

Note that the total variance propagated by regulation flow is $G(s, t) + E(s, t) = (a^2 + e^2 + c^2) \times S(s, t)$. As we standardized all the gene expressions in all our experiments, we have $a^2 + e^2 + c^2 = 1$, so $G(s, t) + E(s, t) = S(s, t)$. $S(s, t)$ is different for each regulation flow, so we cannot directly compute $E(s, t)$ from $G(s, t)$.

3. Experiment and Result

In this section, we performed two experiments. An overview of the data source and experiment information is shown in Figure 3. The pre-processing of the data is discussed at first. The following experiment "Drug target preference" is a significance test for the overlap between genes influenced by drugs and genes driven by strong heritable effects. We found there are more drugs preferring genes driven by strong heritable effects than non-heritable effects.

In the second experiment "Drug influence flows identification", we treat the problem of predicting drug influence flows as a classification problem using heritable, non-heritable and regulation flow strength as the classifier features. Our result shows that heritable flow strength is the best feature for drug influence flow prediction.
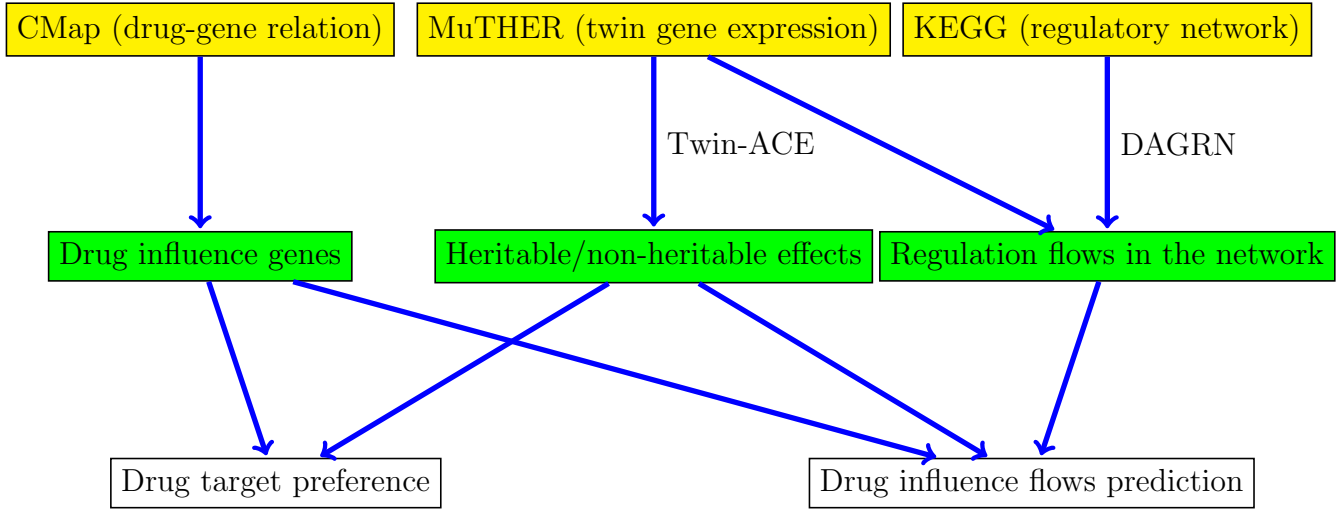
Fig. 3. A process flow chart of the relationship between data and experiments. Yellow nodes are the dataset we used. Green nodes are the information we extracted from the dataset. White nodes are the experiment we conducted.

3.1. *Data pre-process*

CMap database[3,4] is used to extract effective gene-drug relationships. The CMap database maintains a rank matrix for gene's differential expression under influence of 6100 drugs. For each drug d, if gene x's expression fold change is in the 99th percentile, we say that gene x is influenced by drug d.

Furthermore, in a regulation flow $s \to t$, if both gene s and t are influenced by a drug d, we deem the flow $s \to t$ a drug d's influence flow.

Our twin gene expression data is acquired from MuTHER[2] database. The data was measured on blood samples from 276 monozygotic and 442 dizygotic twins. All the gene expression data is centered and standardized. Hence, the expression vector of each gene has mean zero and unit variance. Heritable and non-heritable effects for each gene are estimated using Twin-ACE model.

We combined a joint gene regulatory network (GRN) from 257 human signal pathways in KEGG pathway[14,15] database. We selected the sub-network from GRN that contains genes from MuTHER and converted it into a DAGRN. We sorted the nodes based on their children count. We then removed edges conflicting with this order, that is to say edges pointing from lower ranked nodes to higher ranked nodes under the order. The transformed DAGRN from KEGG pathway contains 28600 directed edges and 3245 genes.

3.2. *Drug target preference*

We test the significance of the overlap between genes influenced by drugs and genes driven by strong heritable effects. 6100 drug instances from CMap database were used for the test. We calculate the heritability of each gene using the ACE parameters as $h = \frac{a^2}{a^2+c^2+e^2}$. Hence, heritability is a value between 0 and 1 indicates how much the percentage of heritable effect in total variance of the gene. The genes with heritability over 0.5 are deemed heritable genes; genes with heritability under 0.2 are deemed non-heritable genes. There are 120 heritable genes and 2199 non-heritable genes. We performed a hypergeometric significance test. There are total $U = 3249$ genes, $M = 120$ of them are heritable. For each drug with $N = 32$ (top 1%) genes influenced, where K of them are heritable genes, we calculate the p-value as the probability of having K or more heritable genes in randomly chosen N samples from total U genes. If p-value is smaller than a certain threshold, we deem the overlap between the genes influenced by the drug and heritable genes significant. We call this kind of drugs "heritable-gene-targeting drugs". We also performed the same significance test to identify "non-heritable-gene-targeting drugs".

The result is shown in Figure 4. When we select p-value threshold as 0.001, there are 4 heritable-gene-targeting drugs and 1 non-heritable-gene-targeting drug. The names of the drugs and the constitution of the genes influenced by the drugs are shown in Figure 5.

Figure 4 shows dramatic difference in counts of heritable-gene-targeting drugs and non-heritable-gene-targeting drugs. Hence, current drugs are more likely to target a gene driven by strong heritable effects rather than strong non-heritable effects.

3.3. *Drug influence flows prediction*

We estimated the regulatory network built from DAGRN using the twin gene expression data from MuTHER database. Stimulate-Response matrix is calculated for 3245 genes in the regulatory network. We selected the threshold $T = 0.3$ to be the threshold for regulation flows and extracted 233 regulation flows. There are 212 different target genes and 164 different source genes. From the CMap, we found 77 regulation flows are true drug influence flows.

To validate our assumption that the regulation flow with high heritable flow strength is more likely to be a drug effective flow, three features are compared here as a classifier of drug influence flow:

- 1. Heritable flow strength of the regulation flow: $G(s,t)$.
- 2. Non-heritable flow strength of the regulation flow: $E(s,t)$.
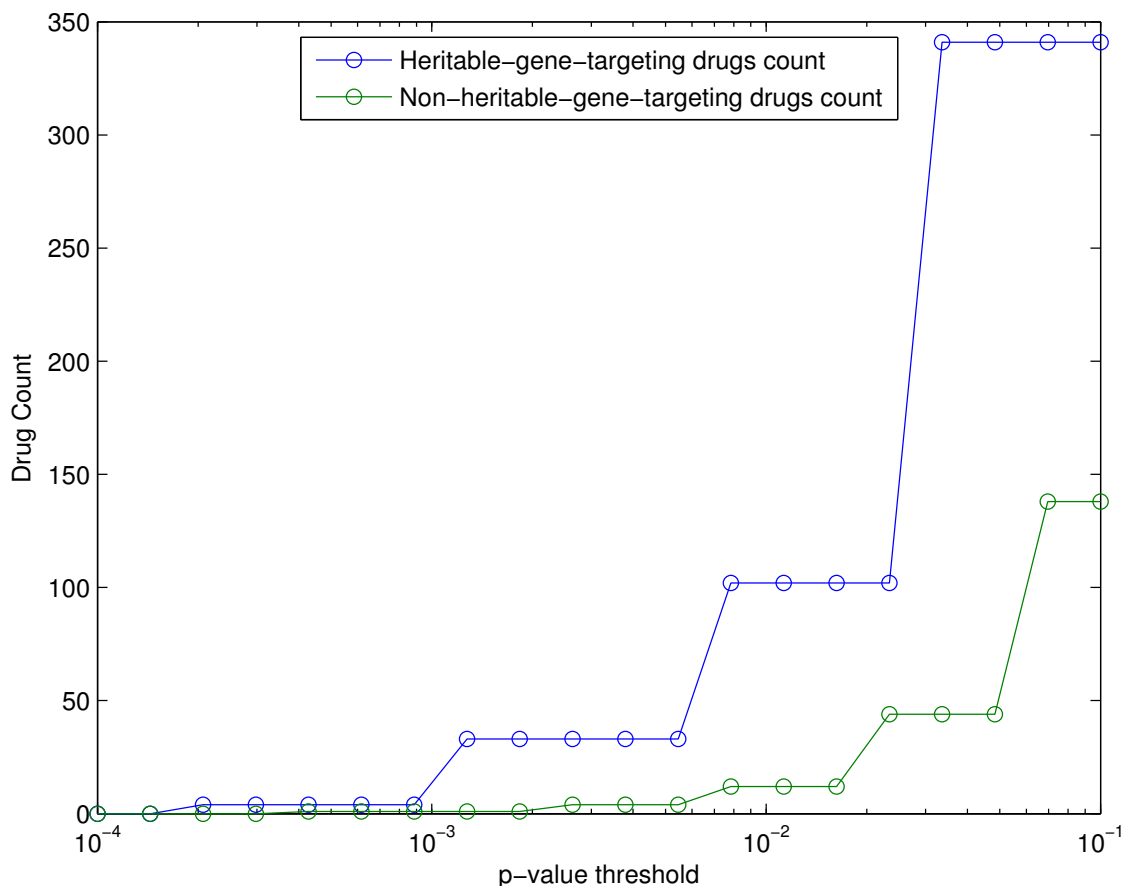- 3. Regulation flow strength of the regulation flow: $S(s,t)$.

Fig. 4. The blue and green lines indicate the counts of drugs that have significant overlap with heritable or non-heritable genes when selecting different p-value thresholds.

For each feature, we selected a list of value thresholds from minimum to maximum of that feature value. If a known drug influence flow has feature value above the threshold, it is counted as a true-positive. The true-positive rate (TPR) is the ratio between true-positive count and total number of drug influence flows. If a regulation flow above the threshold is not a drug influence flow, it is counted as a false-positive. The false-positive rate (FPR) is the ratio between false-positive count and total number of regulation flows that are not drug influence flows. The TPRs and FPRs across all the thresholds construct receiver operating characteristic (ROC) curve. We use area-under-curve (AUC) as the performance metric. A higher AUC indicates the feature is better for predicting correct drug influence flows. The ROC curves for three features are plotted in Figure 6. It is obvious from the result that using the feature heritable flow strength (AUC = 0.63) is much better than non-heritable flow strength (AUC = 0.38) and regulation flow strength (AUC = 0.44) for predicting drug influence flows.

We also listed the drug influence flows ranked by heritable and non-heritable flow strength in Table 1 and 2.
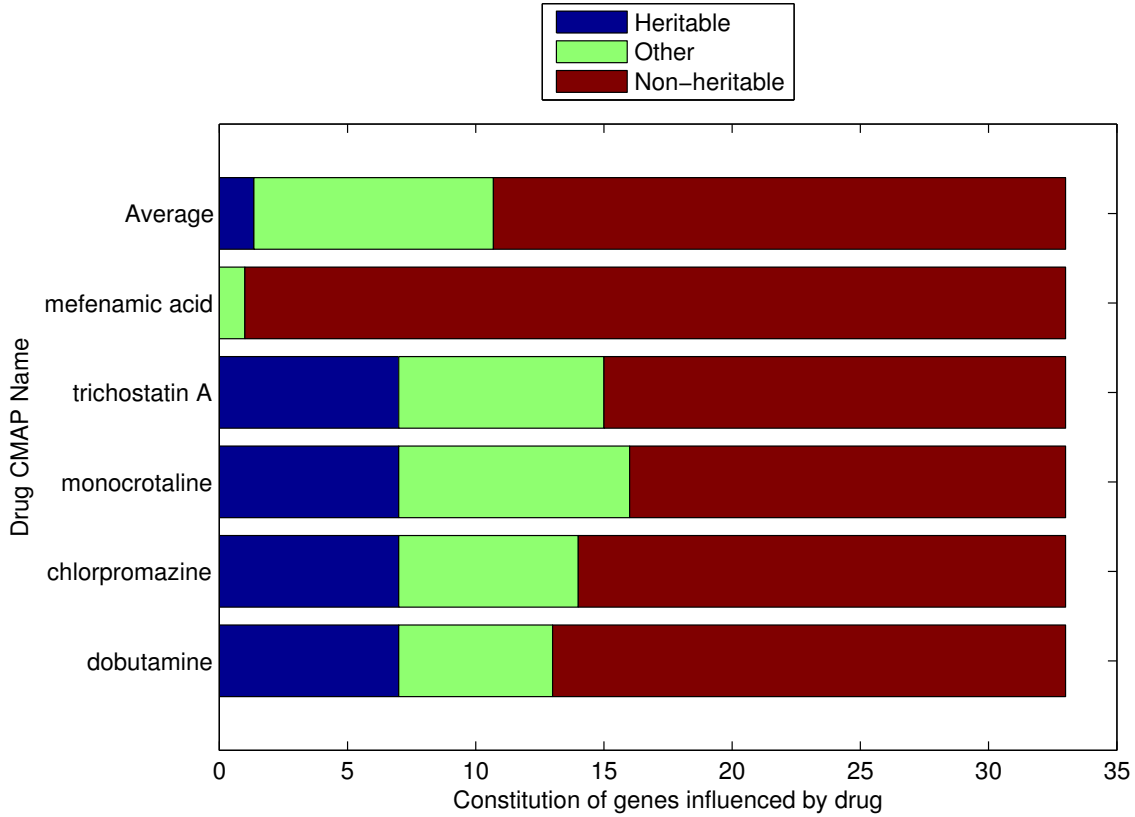
Fig. 5. Non-heritable-gene-targeting drug ("mefenamic acid") and heritable-gene-targeting drugs "richostatin A,monocrotaline chlorpromazine,dobutamine"). The "Average" bar shows the average constitution of all genes influenced by drugs.

Table 1. Drug influence flows ranked by heritable flow strength (top 10)

Target	Source	G(s,t)	E(s,t)	drug CMap name
ATG12	FOXO3	0.43	0.33	deferoxamine,famotidine,Prestwick-860,azacitidine,bupivacaine
ARPC1B	ACTG1	0.34	0.24	mesoridazine
INADL	CLDN15	0.30	0.29	piroxicam
ARPC4	ACTG1	0.28	0.20	benzylpenicillin,zardaverine
ARPC3	ACTG1	0.25	0.18	phenazopyridine
DGUOK	GUK1	0.24	0.29	benperidol
BCL2L11	DDIT3	0.20	0.21	ChicagoSkyBlue6B
CD22	PTPN6	0.19	0.26	haloperidol,6-bromoindirubin-3'-oxime
TUBB	TUBA1C	0.18	0.65	PF-00539758-00
TK1	DUT	0.18	0.63	tanespimycin

4. Discussion

In this paper, we answered the very first question in our paper: whether a drug aiming to perturb a disease phenotype should target genes whose expression is dominated by heritable or non-heritable factors?

Even though the variance in expression of most genes in the blood sample are driven by non-heritable effects, our first experiment showed that drugs prefer to influence genes driven
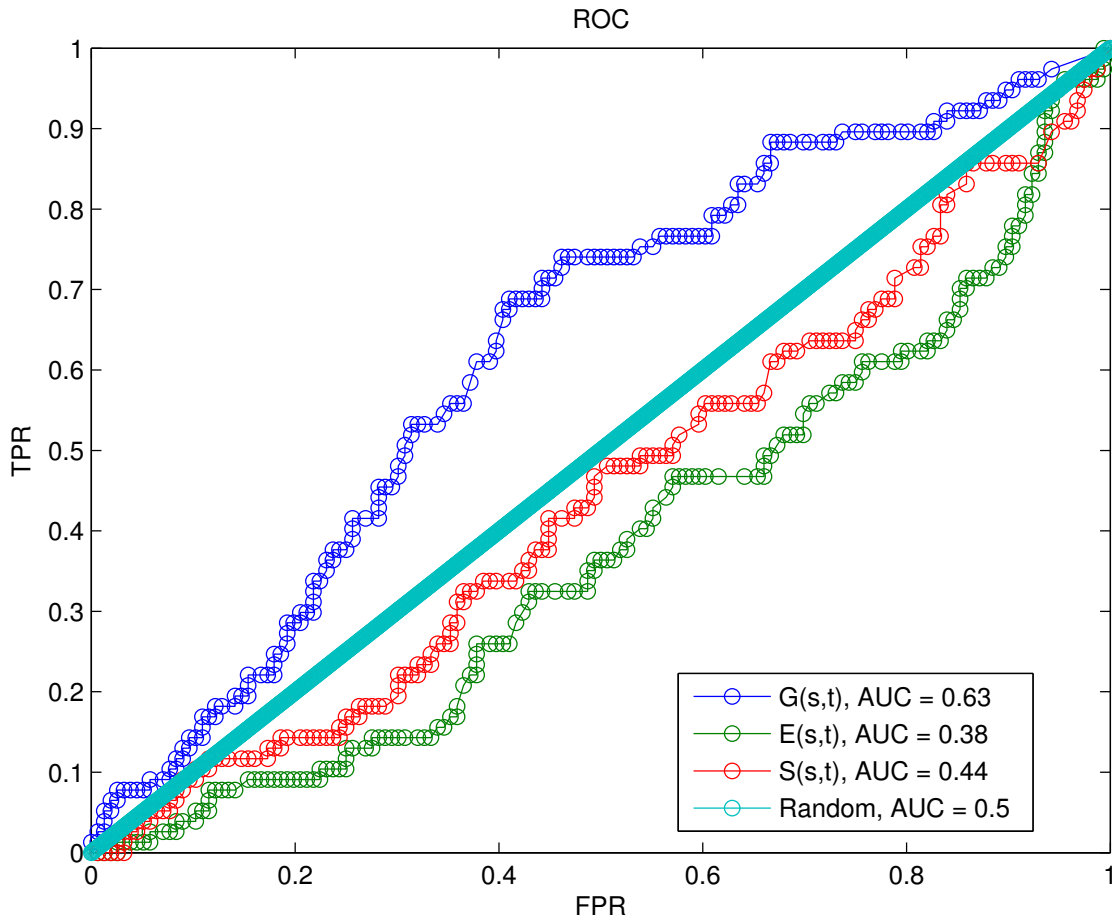
Fig. 6. The receiver operating characteristic curve and AUC for three features as classifiers for drug influence flows.

Table 2. Drug influence flows ranked by non-heritable flow strength (top 10)

Target	Source	G(s,t)	E(s,t)	drug CMap name
SUCLA2	SDHD	0.09	0.75	epirizole
ANAPC10	MAD2L1	0.04	0.67	nomifensine
TUBB	TUBA1C	0.18	0.65	PF-00539758-00
BTG3	PABPC1	0.03	0.64	deptropine,phenoxybenzamine,oxprenolol
TK1	DUT	0.18	0.63	tanespimycin
ENOPH1	APIP	0.10	0.62	arcaine
INPP5E	PLCB2	0.12	0.60	furosemide
PFAS	GART	0.13	0.55	S-propranolol
MAP3K4	GADD45B	0.03	0.52	trichostatinA
IVD	HADHA	0.12	0.51	alfuzosin,lomefloxacin,isometheptene,sulfaquinoxaline

by strong heritable effects rather than non-heritable effects. We then extended this observation to the in the background of regulatory network, where we found the regulation flow with high heritable flow strength is more likely to be a drug influence flow than flows with high non-heritable strength. The answer to the question is clear from our experiment: a drug aiming to

perturb a disease phenotype should target genes whose expression is dominated by heritable rather than non-heritable factors.

In both experiments, we identified the drugs targeting genes driven by strong heritable or non-heritable effects. These observations and discoveries can help us design drugs targeting more specific and precise regulation flows in the regulatory network to influence the final target gene's expression.

There are plenty of extensions for the current method. We can remove the DAGRN restriction and construct a model the general gene regulatory network with loops and undirected edges, as this is the most common gene regulation pathway. Another possible application is to use the Stimulation-Response matrix for any specific study in gene expression control problem with more constrained goals and resource limitations.

Acknowledgments

References

1. E. Klechevsky, R. Morita, M. Liu, Y. Cao, S. Coquery, L. Thompson-Snipes, F. Briere, D. Chaussabel, G. Zurawski, A. K. Palucka *et al.*, *Immunity* **29**, 497 (2008).
2. A. C. Nica, L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, K. Small *et al.*, *PLoS genetics* **7**, p. e1002003 (2011).
3. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, *science* **313**, 1929 (2006).
4. J. Lamb, *Nature Reviews Cancer* **7**, 54 (2007).
5. E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett *et al.*, *Nature genetics* **44**, 1084 (2012).
6. W. Sadee, *Clinical Pharmacology & Therapeutics* **92**, 428 (2012).
7. C. J. Patel and M. R. Cullen, Genetic variability in molecular responses to chemical exposure, in *Molecular, Clinical and Environmental Toxicology*, (Springer, 2012) pp. 437–457.
8. C. Cotsapas (2008).
9. M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási and M. Vidal, *Nature biotechnology* **25**, 1119 (2007).
10. P. Lecca and C. Priami, *Drug discovery today* **18**, 256 (2013).
11. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006).
12. F. V. Rijsdijk and P. C. Sham, *Briefings in Bioinformatics* **3**, 119 (2002).
13. J. Nocedal and S. J. Wright, Quasi-newton methods, in *Numerical Optimization*, (Springer Series in Operations Research and Financial Engineering, 2006).
14. M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe, *Nucleic acids research* **42**, D199 (2014).
15. M. Kanehisa and S. Goto, *Nucleic acids research* **28**, 27 (2000).