

## METHODOLOGY ARTICLE

## Open Access

# Estimating causal effects with a *non-paranormal* method for the design of efficient intervention experiments

Reiji Teramoto<sup>\*</sup>, Chiaki Saito and Shin-ichi Funahashi

## Abstract

**Background:** Knockdown or overexpression of genes is widely used to identify genes that play important roles in many aspects of cellular functions and phenotypes. Because next-generation sequencing generates high-throughput data that allow us to detect genes, it is important to identify genes that drive functional and phenotypic changes of cells. However, conventional methods rely heavily on the assumption of normality and they often give incorrect results when the assumption is not true. To relax the Gaussian assumption in causal inference, we introduce the *non-paranormal* method to test conditional independence in the PC-algorithm. Then, we present the *non-paranormal* intervention-calculus when the directed acyclic graph (DAG) is absent (NPN-IDA), which incorporates the cumulative nature of effects through a cascaded pathway via causal inference for ranking causal genes against a phenotype with the *non-paranormal* method for estimating DAGs.

**Results:** We demonstrate that causal inference with the *non-paranormal* method significantly improves the performance in estimating DAGs on synthetic data in comparison with the original PC-algorithm. Moreover, we show that NPN-IDA outperforms the conventional methods in exploring regulators of the flowering time in *Arabidopsis thaliana* and regulators that control the browning of white adipocytes in mice. Our results show that performance improvement in estimating DAGs contributes to an accurate estimation of causal effects.

**Conclusions:** Although the simplest alternative procedure was used, our proposed method enables us to design efficient intervention experiments and can be applied to a wide range of research purposes, including drug discovery, because of its generality.

**Keywords:** Non-paranormal, Gaussian assumption, Causal effect, Intervention-calculus, Directed acyclic graph, Machine learning, Causal inference, Experiment design

## Background

Intervention experiments, e.g., knockdown or overexpression, are commonly conducted to identify genes that determine cell fates such as differentiation [1], induction of pluripotency [2], and direct reprogramming [3]. Those experiments are now indispensable in biological and medical research. Although intervention experiments identify a causal gene responsible for a phenotype of interest, they are time-consuming and cost-intensive. Therefore, it is very important to prioritize and focus on causal genes with high confidence. However, it is difficult to infer causal effects

only from observational data. This task coincides with inferring causal effects that are established in Statistics. Note that in this problem setting, a causal effect is different from a regression-type effect of association [4]. In fact, previous studies suggested that representative regression methods such as lasso and elastic net are not appropriate for our goal [4-6].

Recently, there has been much progress to address this problem by employing the intervention-calculus when the directed acyclic graph (DAG) is absent (IDA) [4-6] for the design of efficient intervention experiments. IDA combines two methods: (1) inference the unknown underlying DAG model from observational data by the PC-algorithm [7] and (2) estimating causal effects on the basis of DAG using intervention-calculus; furthermore, it provides estimated

\* Correspondence: [teramoto@forerunner-pharma.co.jp](mailto:teramoto@forerunner-pharma.co.jp)  
Department for Research, Forerunner Pharma Research, Co., Ltd, Yokohama, Japan

lower bounds of total causal effects from observational data. The PC-algorithm is computationally feasible and appropriate for high-dimensional settings. In addition, it has the desirable property to achieve high computational efficiency as a function of sparseness of the true underlying DAG model.

In spite of these advantages, the PC-algorithm requires the Gaussianity assumption to use partial correlations to test conditional independence, and this assumption is not necessarily true in real data sets. Because the normality assumption is restrictive and conclusions inferred under this assumption could be misleading, it is desirable to relax the Gaussian assumption.

On the other hand, *non-paranormal* methods that use a semiparametric Gaussian copula have been proposed for estimating sparse undirected graphs and exhibit significant improvement in the performance because the normality assumption is relaxed [8,9]. The main idea of the *non-paranormal* method is to exploit the nonparametric correlation coefficient instead of Pearson's correlation coefficient for estimation. Although this is the simplest alternative procedure, the *non-paranormal* graphical model could be a viable alternative for the Gaussian graphical model.

Consequently, we present *non-paranormal* IDA (NPN-IDA), which uses nonparametric partial correlations to test conditional independencies in the PC-algorithm for intervention-calculus. In our method, the Gaussian assumption in the PC-algorithm is naturally relaxed by using nonparametric partial correlation. Although the *non-paranormal* method has been successfully applied to estimating undirected graphs in previous studies, we show that it works well for estimating DAGs in the PC-algorithm. Next, we applied our method to *Arabidopsis thaliana* microarray data and mouse microarray data to demonstrate that NPN-IDA outperforms IDA in exploring regulators of the flowering time in *A. thaliana* and regulators that control the browning of white adipocytes in mice.

In summary, the three main contributions of this work are: (1) introduction of a *non-paranormal* method for inference of the unknown underlying DAG model from observational data in the expansive framework of the PC-algorithm, (2) combination of the *non-paranormal* method and the PC-algorithm significantly improves the performance in estimating DAGs on synthetic data, and (3) NPN-IDA is effective in exploring regulators that control specific phenotypes of interest.

## Methods

We first introduce the IDA procedure. IDA consists of (1) inference of the unknown underlying DAG model from observational data by PC-algorithm and (2) estimation of causal effects based on the DAG using intervention-

calculus. Then, we introduce the *non-paranormal* method for PC-algorithm. Finally, we present the combination of the *non-paranormal* method for PC-algorithm and estimating causal effects as NPN-IDA algorithm.

### Inference DAGs with the PC-algorithm

Let  $G = (V, E)$  be a graph consisting of vertices  $V$  and a set of edges  $E \subseteq V \times V$ . In our context, the vertices represent random variables  $X_1, \dots, X_p$ , and  $Y$ . The edges represent relationships between pairs of these variables. It is possible that some DAGs fulfill the Markov condition and encode the same list of conditional independencies. Two DAGs are observationally equivalent only if they have the same skeleton and same sets of v-structures, i.e., two converging arrows whose tails are connected by an arrow. In this way, DAGs can be partitioned into equivalent classes, where all members are observationally equivalent and represent the same conditional independence. In a given conditional independence set of DAGs, one can only determine a DAG up to its equivalence class. The equivalence class is called completed partially directed acyclic graph (CPDAG). It has the same skeleton as every DAG in the equivalence class and directed edges only where all DAGs in the equivalence class have the same directed edge. Arrows that point into one direction for some DAGs in the equivalence class and in the other direction for other DAGs in the equivalence class are represented by undirected edges (Figure 1). By assuming that random variables are multivariate normally distributed, conditional independencies can be inferred from a partial correlation between  $X_i$  and  $X_j$  given a set of other variables  $S$  that equals zero:

$$\rho_{ij|S} = \text{Parcor}(X_i \perp X_j | X_S) = 0 \quad \text{iff} \quad X_i \perp X_j | X_S \quad (1)$$

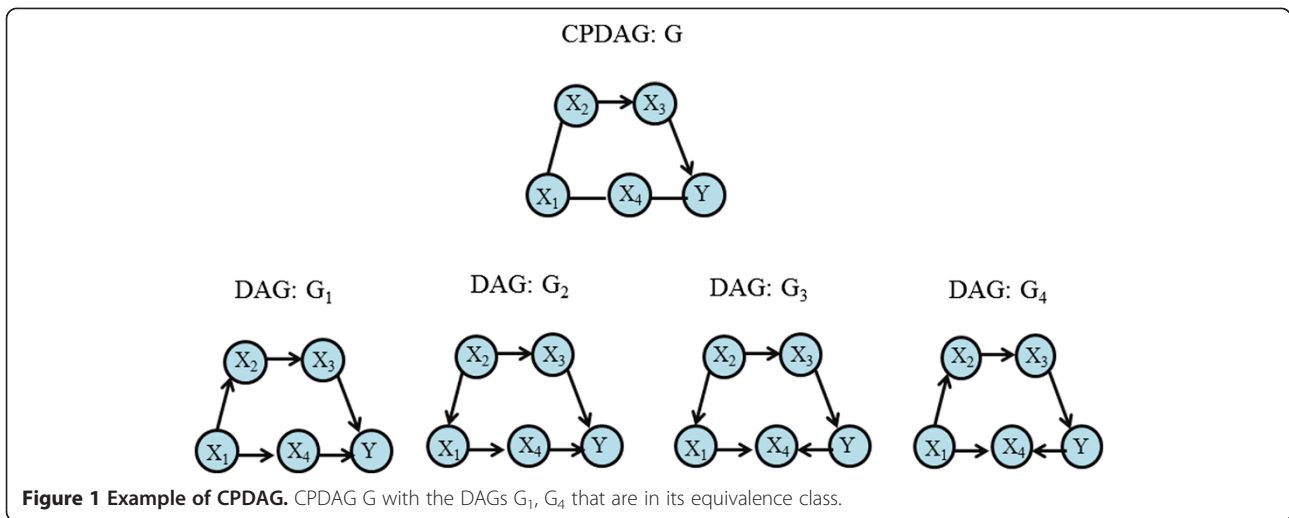
We then used the sample version of the PC-algorithm to estimate the corresponding CPDAG. This involves multiple testing for Fisher's Z-transformed partial correlations,

$$\hat{Z}_{ij|S} = \frac{1}{2} \log \left( \frac{1 + \hat{\rho}_{ij|S}}{1 - \hat{\rho}_{ij|S}} \right).$$

Because  $\hat{Z}_{ij|S}$  has a  $N(0, (n - |S| - 3)^{-1})$  distribution if  $\rho_{ij|S} = 0$ , we conclude that  $\rho_{ij|S} \neq 0$  if

$$|\hat{Z}_{ij|S}| \sqrt{n - |S| - 3} > \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right),$$

where  $\Phi$  is the standard normal distribution function and  $\alpha$  is a tuning parameter, which can be interpreted as the significance level of a single partial correlation test. Choosing an appropriate value for  $\alpha$  is difficult but, for example, can be done with the Bayesian information criterion.



First, the PC-algorithm generates a skeleton on the basis of conditional independencies. The outline of the PC-algorithm is shown in Figure 2. The complete PC-algorithm is described in detail in a previous work [7]. Note that the PC-algorithm employs partial correlation to test conditional independency.

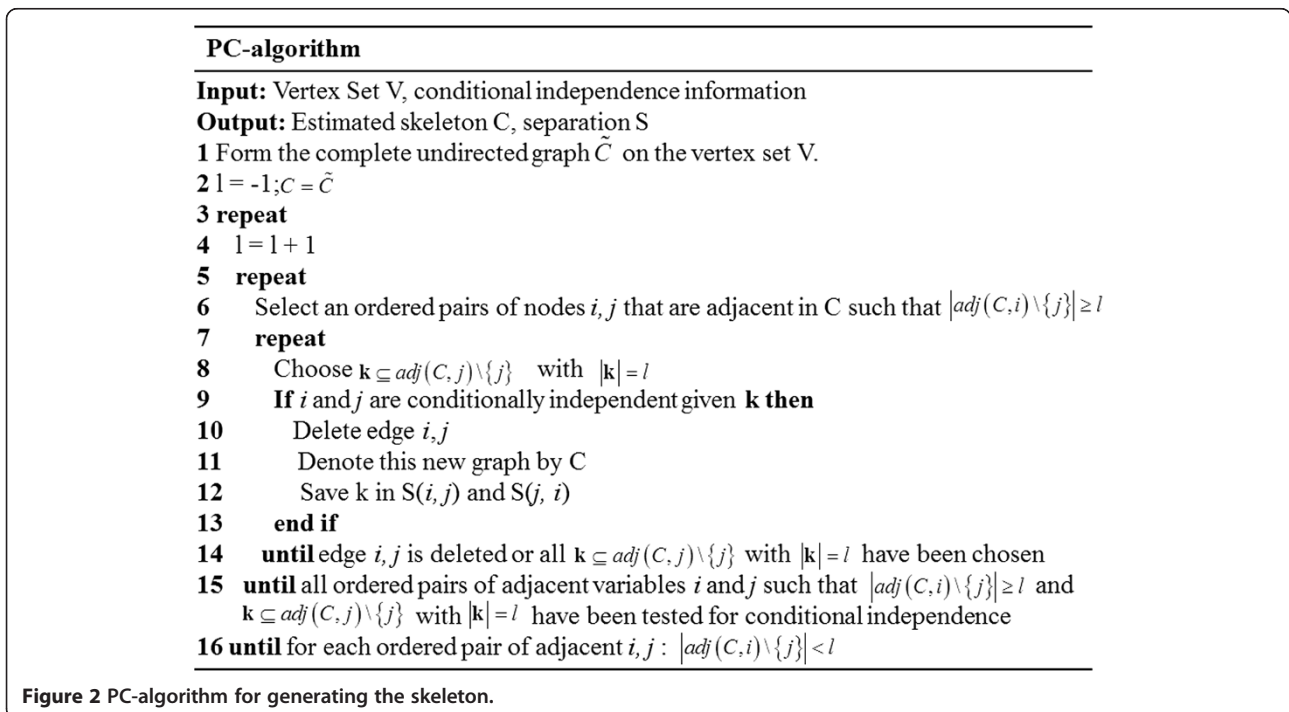
#### Estimating causal effects using intervention-calculus

Again, we considered  $p + 1$  random variables  $X_1, \dots, X_p, Y$  (also referred to as  $X_1, \dots, X_p, X_{p+1}$ ). Any distribution that is generated from a DAG with independent error is

called Markovian with respect to the DAG. Therefore, any Markovian distribution can be factorized as

$$f(x_1, \dots, x_{p+1}) = \prod_{j=1}^{p+1} f(x_j | pa_j).$$

To represent the effect of an intervention of a set of variables, a *do* operator is introduced. We denote the distribution of  $Y$  that would occur if the treatment condition  $X_i = x'_i$  was enforced uniformly over the population via some intervention as  $f(y | do(X_i = x'_i))$ . For a Markovian



**IDA algorithm**

**Input:** CPDAG  $G$ , conditional independencies of  $Y, X_1, \dots, X_p$   
**Output:** Matrix  $\Theta$  of possible causal effects  
**1** Determine all DAGs  $G_1, \dots, G_m$  in the equivalence class  
**2** for  $j = 1$  to  $m$  do  
**3** for  $i = 1$  to  $p$  do  
**4**  $\theta_{ij} = \beta_{i,pa_i(G_j)}$   
**5** end  
**6** end

Figure 3 IDA algorithm.

model, the distribution generated by an intervention  $do(X_i = x'_i)$  on the set of variables  $X_1, \dots, X_{p+1}$  is given by the following truncated factorization formula:

$$f(x_1, \dots, x_{p+1} | do(X_i = x'_i)) = \begin{cases} \prod_{j=1, j \neq i}^{p+1} f(x_j | pa_j) |_{x_i=x'_i}, & \text{if } x_i = x'_i \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

Where  $f(x_j | pa_j)$  are the pre-intervention conditional distributions. Note that this formula employs the DAG structure to write the interventional distribution on the left-hand side in terms of pre-intervention conditional distributions on the right-hand side. The distribution of  $Y = X_{p+1}$  after an intervention  $do(X_i = x'_i)$  can be

obtained by marginalizing out  $x_1, \dots, x_p$  in equation (2). This can be written as follows:

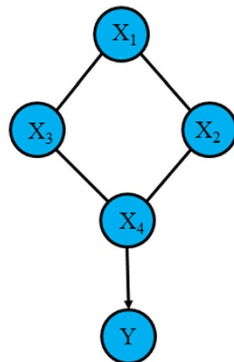
$$f(y | do(X_i = x'_i)) = \begin{cases} f(y), & \text{if } Y \in pa_i, \\ \int f(y | x'_i, pa_i) f(pa_i) dpa_i & \text{if } Y \notin pa_i, \end{cases}$$

where  $f(\cdot)$  and  $f(\frac{\ddot{A}}{n} | x'_i, pa_i)$  represent pre-intervention distributions. We can summarize the distribution generated by the intervention by its mean

$$E(Y | do(X_i = x'_i)) = \begin{cases} E(Y), & \text{if } Y \in pa_i \\ \int E(Y | x'_i, pa_i) f(pa_i) dpa_i, & \text{if } Y \notin pa_i \end{cases}$$

and define the causal effect of  $do(X_i = x'_i)$  on  $Y$  by

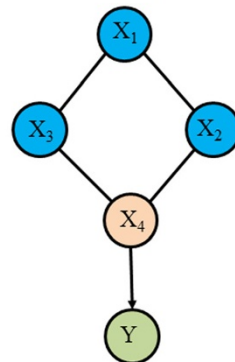
If not all edges can be directed, we have a completed partially directed acyclic graph (CPDAG).



PC-algorithm

Estimate causal effect of a given gene  $X_j$  on a target variable  $Y$  as regression coefficient  $\beta_j$  of the linear regression.

$$Y = \beta_j X_j + \beta_{pa(X_j)} pa(X_j) + \varepsilon$$



Intervention-calculus

Figure 4 Illustrative procedure of IDA.

### NPN-IDA algorithm

**Input:** Variables  $Y, X_1, \dots, X_p \in \mathbb{R}^n$   
**Output:** Matrix  $\Theta$  of possible causal effects

- 1 Compute Spearman's correlation coefficient  $\mathbf{r}^s \in \mathbb{R}^{p \times p}$
- 2 Estimate *non-paranormal* correlation coefficient  $\Sigma = 2 \sin\left(\frac{\pi}{6} \mathbf{r}^s\right)$
- 3 Run the PC-algorithm for estimating a CPDAG  $G$
- 4 Determine all DAGs  $G_1, \dots, G_m$  in the equivalence class
- 5 **for**  $j = 1$  to  $m$  **do**
- 6   **for**  $i = 1$  to  $p$  **do**
- 7      $\theta_{ij} = \beta_{i|pa_i(G_j)}$
- 8   **end**
- 9 **end**

Figure 5 NPN-IDA algorithm.

$$\frac{\partial}{\partial x} E(Y | do(X_i = x)) \Big|_{x=x'_i} \quad (3) \quad \int E(Y | x'_i, pa_i) f(pa_i) dpa_i = \gamma_i x'_i + \int \gamma_{pa_i}^T pa_i f(pa_i) dpa_i,$$

Under the assumption that  $X_1, \dots, X_p, Y$  are jointly Gaussian, it is easy to compute the causal effects. Gaussianity implies that  $E(Y | x'_i, pa_i)$  is linear in  $x'_i$  and  $pa_i$ . That is

$$E(Y | x'_i, pa_i) = \gamma_0 + \gamma_i x'_i + \gamma_{pa_i}^T pa_i$$

for some values,  $\gamma_0$ ,  $\gamma_i$  and  $\gamma_{pa_i} \in \mathbb{R}^{|pa_i|}$ , where  $|pa_i|$  is the cardinality of the set  $pa_i$ . Then, we derive

which is linear in  $x'_i$ . The causal effect of  $X_i$  on  $Y$  when  $Y \notin pa_i$  can be computed with (3) yielding  $\gamma_i$ , which is simply the regression coefficient of  $X_i$  in the regression of  $Y$  on  $X_i$  and  $pa_i$ . When  $Y \in pa_i$ , the causal effect becomes zero, because  $Y$  is a direct cause of  $X_i$ . Therefore, the causal effect of  $X_i$  on  $Y$  is given by the following equation:

$$\beta_{i|pa_i} = \begin{cases} 0, & \text{if } Y \in pa_i \\ \text{coefficient of } X_i \text{ in } Y \sim X_i + pa_i, & \text{if } Y \notin pa_i \end{cases}$$

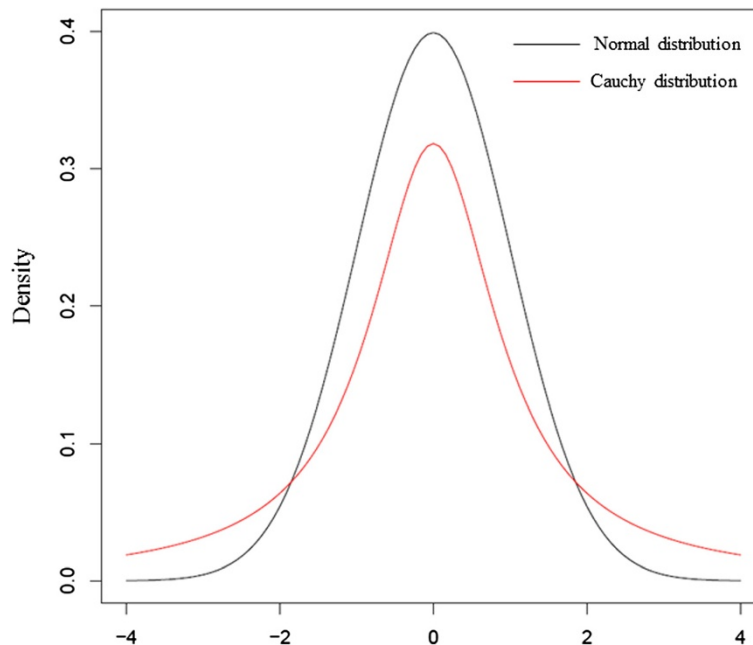


Figure 6 Comparison of standard normal distribution and standard Cauchy distribution.

**Table 1 Parameter setting for performance evaluation of the NPN-PC algorithm**

Parameters	Set values
$\alpha$	$10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$
$p$	100, 200
$n$	50, 100
$cp$	0.005, 0.01
Mixing rate	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1

$\alpha$ : significant level,  $p$ : number of variables,  $n$ : number of samples,  $cp$ : probability of connecting one node to another node.

where  $Y \sim X_i + pa_i$  is a shorthand notation for the linear regression of  $Y$  on  $X_i$  and  $pa_i$ . Consequently, in the Gaussian case, the causal effect does not depend on the value of  $x'_i$  and can be interpreted as

$$\beta_{i|pa_i} = E\left[Y|do(X_i = x'_i + 1)\right] - E\left[Y|do(X_i = x'_i)\right]$$

for any value of  $x'_i$ .

Next, we will describe the IDA algorithm in detail. For simplicity, we only show the case in which the PC-algorithm gives the correct CPDAG. Details of the sample version of the IDA algorithm can be found in previously conducted studies [5-7]. On the basis of CPDAG, which is inferred from the PC-algorithm, we can compute the causal effect for every DAG in the equivalence class, which consists of multisets. Multisets differ from normal sets in that the multiplicity of the elements is taken into account. The IDA algorithm is given in Figure 3.

For computing the causal effects  $\theta_{ij}$  of  $X_i$  on  $Y$  in DAG  $G_p$ , the IDA algorithm simply computes the regression coefficient of  $X_i$  in a regression of  $Y$  on  $X_i$  and  $pa_i$ , which is denoted by  $\beta_{i|pa_i(G_i)}$ . This procedure is performed for every DAG in the equivalence class yielding a vector of length  $j$  of possible causal effects, where  $j$  is the number of DAGs in the equivalence class. Computing the effect of every  $X_1, \dots, X_p$  on  $Y$  yields a matrix  $\Theta$  with  $\theta_{ij}$  entries. We describe the illustrative procedure of IDA in Figure 4. When we obtain a single value for the estimated causal effect, we can take the minimal absolute value of the multiset as a lower bound from  $\Theta$  for the true absolute intervention effect. This procedure intends to reduce the number of false positives. From a practical point of view, because the number of false positive should be kept as low as possible, it is desirable to provide conservative results.

#### Non-paranormal method for NPN-IDA

In the PC-algorithm, the conditional independency is inferred from a sample partial correlation  $\hat{\rho}_{ij|S}$  between  $X_i$  and  $X_j$  given a set of other variables  $S$ . As described in the section of the PC-algorithm, a sample partial correlation coefficient can be obtained by calculating a sample

Pearson's correlation coefficient. This fact enables us to relax the Gaussianity assumption of the PC-algorithm by using the *non-paranormal* method. On the basis of [9], we derive the *non-paranormal* version of the PC-algorithm (NPN-PC algorithm). Let  $f = (f_1, \dots, f_p)$  be a set of monotonic univariate functions and let  $\Sigma^0 \in \mathbb{R}^{p \times p}$  be a positive definite correlation matrix with  $diag(\Sigma^0) = 1$ . We suppose that the  $p$ -dimensional random variable  $X = (X_1, \dots, X_p)^T$  has a *non-paranormal* distribution  $X \sim NPN_p(f, \Sigma^0)$  if  $f(X) \equiv (f_1(X_1), \dots, f_p(X_p))^T \sim N(0, \Sigma^0)$ . For continuous distributions, the *non-paranormal* family is equivalent to the Gaussian copula family [8]. It has been shown that the *non-paranormal* family is much richer than the normal family. The conditional independence is encoded by  $\Omega^0 = (\Sigma^0)^{-1}$ . Therefore, we can write the following formula given a set of other variables  $S$ :

$$\Omega_{ij}^0 = \rho_{ij|S} = 0 \text{ iff } X_i \perp\!\!\!\perp X_j | X_S \quad (4)$$

For Gaussian copula distributions, the following important lemma connects Spearman's correlation coefficient  $r_{ij}^s$  to the underlying Pearson's correlation coefficient [8,9].

**Lemma 1.** [10] By assuming  $X \sim NPN(f, \Sigma^0)$ , we have  $\Sigma_{ij}^0 = 2 \sin\left(\frac{\pi}{6} r_{ij}^s\right)$ .

According to the lemma, we can define the following estimator  $\hat{S}^\rho = [\hat{S}_{ij}^\rho]$  for the unknown correlation matrix  $\Sigma^0$ :  $\hat{S}_{ij}^\rho = 2 \sin\left(\frac{\pi}{6} \hat{r}_{ij}^s\right)$  for  $i \neq j$ , and  $diag(\hat{S}^\rho) = 1$ . Finally, if we define  $p = (\hat{S}^\rho)^{-1}$ , we obtain the following formula that connects Spearman's correlation coefficient and the *non-paranormal* partial correlation coefficient  $\rho_{ij|S^s}$ .

$$\rho_{ij|S^s} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}$$

Instead of  $\rho_{ij|S}$  in (1), we employ  $\rho_{ij|S^s}$  to test conditional independence for estimating CPDAG in the PC-algorithm. For simplicity, we denoted the method that combines the NPN-PC algorithm and IDA as NPN-IDA. We describe the pseudo code of the NPN-IDA algorithm in Figure 5.

## Results

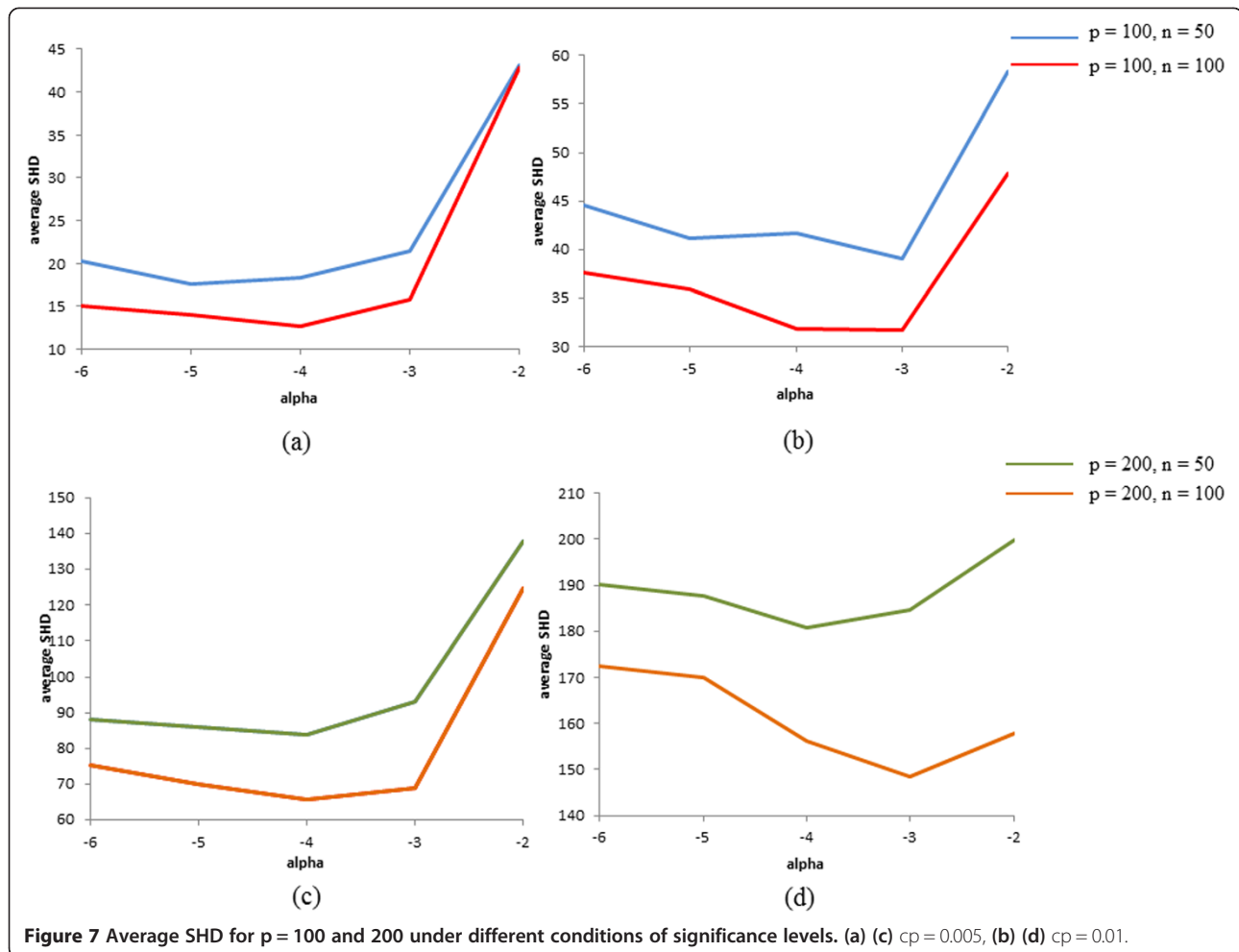
### Experimental settings

Two experiments were conducted for different purposes. The first purpose was to evaluate the performance of the NPN-PC algorithm on synthetic data when the Gaussianity assumption is not true. According to a previous study [7], the four metrics of performance, i.e., true positive rate (TPR), false positive rate (FPR), true discovery rate (TDR), and structural hamming distance (SHD), are used. TPR, FPR, and TDR are defined as follows:

**Table 2 Performance comparison between PC-algorithm and NPN-PC algorithm**

p	n	Mixing rate	TPR		FPR		TDR		SHD	
			PC	NPN-PC	PC	NPN-PC	PC	NPN-PC	PC	NPN-PC
100	50	0.1	0.360	<b>0.373</b>	0.0021	<b>0.0003</b>	0.468	<b>0.876</b>	28.72	<b>19.40</b>
		0.2	0.341	<b>0.395</b>	0.0035	<b>0.0002</b>	0.326	<b>0.899</b>	35.74	<b>18.42</b>
		0.3	0.316	<b>0.372</b>	0.0044	<b>0.0002</b>	0.263	<b>0.898</b>	40.08	<b>18.10</b>
		0.4	0.325	<b>0.387</b>	0.0046	<b>0.0003</b>	0.266	<b>0.886</b>	42.58	<b>19.88</b>
		0.5	0.308	<b>0.393</b>	0.0047	<b>0.0002</b>	0.257	<b>0.908</b>	43.30	<b>18.92</b>
		0.6	0.326	<b>0.399</b>	0.0049	<b>0.0002</b>	0.252	<b>0.890</b>	42.88	<b>18.68</b>
		0.7	0.310	<b>0.409</b>	0.0051	<b>0.0002</b>	0.238	<b>0.893</b>	45.14	<b>18.80</b>
		0.8	0.297	<b>0.405</b>	0.0052	<b>0.0002</b>	0.221	<b>0.898</b>	44.64	<b>18.04</b>
		0.9	0.319	<b>0.432</b>	0.0052	<b>0.0002</b>	0.237	<b>0.903</b>	44.64	<b>17.74</b>
	1	0.313	<b>0.416</b>	0.0051	<b>0.0002</b>	0.241	<b>0.904</b>	44.74	<b>18.24</b>	
	100	0.1	0.487	<b>0.625</b>	0.0031	<b>0.0001</b>	0.440	<b>0.956</b>	31.38	<b>13.94</b>
		0.2	0.451	<b>0.623</b>	0.0048	<b>0.0002</b>	0.331	<b>0.954</b>	42.30	<b>15.04</b>
		0.3	0.457	<b>0.653</b>	0.0058	<b>0.0001</b>	0.282	<b>0.963</b>	46.04	<b>13.60</b>
		0.4	0.442	<b>0.643</b>	0.0063	<b>0.0002</b>	0.265	<b>0.942</b>	48.84	<b>14.14</b>
		0.5	0.425	<b>0.667</b>	0.0064	<b>0.0001</b>	0.249	<b>0.965</b>	49.50	<b>12.90</b>
		0.6	0.449	<b>0.642</b>	0.0064	<b>0.0002</b>	0.264	<b>0.950</b>	49.32	<b>13.92</b>
		0.7	0.454	<b>0.681</b>	0.0064	<b>0.0002</b>	0.274	<b>0.951</b>	50.16	<b>13.50</b>
		0.8	0.421	<b>0.665</b>	0.0064	<b>0.0002</b>	0.257	<b>0.957</b>	49.78	<b>13.68</b>
		0.9	0.437	<b>0.692</b>	0.0064	<b>0.0001</b>	0.254	<b>0.965</b>	48.80	<b>12.30</b>
1	0.401	<b>0.672</b>	0.0064	<b>0.0002</b>	0.239	<b>0.950</b>	50.00	<b>12.66</b>		
200	50	0.1	0.291	<b>0.315</b>	0.0011	<b>0.0002</b>	0.564	<b>0.894</b>	105.72	<b>86.50</b>
		0.2	0.268	<b>0.326</b>	0.0018	<b>0.0002</b>	0.430	<b>0.892</b>	120.72	<b>85.08</b>
		0.3	0.256	<b>0.323</b>	0.0021	<b>0.0002</b>	0.386	<b>0.906</b>	127.88	<b>85.28</b>
		0.4	0.257	<b>0.317</b>	0.0023	<b>0.0002</b>	0.366	<b>0.903</b>	130.22	<b>84.86</b>
		0.5	0.255	<b>0.324</b>	0.0023	<b>0.0002</b>	0.360	<b>0.897</b>	132.10	<b>86.32</b>
		0.6	0.258	<b>0.345</b>	0.0024	<b>0.0002</b>	0.347	<b>0.885</b>	129.06	<b>81.48</b>
		0.7	0.249	<b>0.335</b>	0.0024	<b>0.0002</b>	0.345	<b>0.908</b>	131.80	<b>82.98</b>
		0.8	0.243	<b>0.335</b>	0.0024	<b>0.0002</b>	0.343	<b>0.888</b>	134.52	<b>85.38</b>
		0.9	0.250	<b>0.338</b>	0.0024	<b>0.0002</b>	0.344	<b>0.896</b>	133.46	<b>83.98</b>
	1	0.249	<b>0.340</b>	0.0024	<b>0.0002</b>	0.350	<b>0.903</b>	133.42	<b>83.94</b>	
	100	0.1	0.424	<b>0.542</b>	0.0018	<b>0.0001</b>	0.534	<b>0.956</b>	110.60	<b>68.40</b>
		0.2	0.387	<b>0.565</b>	0.0027	<b>0.0001</b>	0.415	<b>0.968</b>	131.66	<b>66.50</b>
		0.3	0.374	<b>0.558</b>	0.0031	<b>0.0001</b>	0.377	<b>0.964</b>	140.38	<b>68.48</b>
		0.4	0.365	<b>0.567</b>	0.0032	<b>0.0001</b>	0.368	<b>0.968</b>	145.72	<b>68.72</b>
		0.5	0.372	<b>0.580</b>	0.0033	<b>0.0001</b>	0.357	<b>0.963</b>	142.34	<b>64.52</b>
		0.6	0.360	<b>0.576</b>	0.0034	<b>0.0001</b>	0.349	<b>0.967</b>	146.92	<b>66.90</b>
		0.7	0.363	<b>0.577</b>	0.0033	<b>0.0001</b>	0.354	<b>0.965</b>	146.84	<b>67.20</b>
		0.8	0.363	<b>0.581</b>	0.0034	<b>0.0001</b>	0.346	<b>0.968</b>	146.18	<b>65.90</b>
		0.9	0.372	<b>0.585</b>	0.0034	<b>0.0001</b>	0.361	<b>0.964</b>	145.62	<b>67.04</b>
1		0.369	<b>0.594</b>	0.0034	<b>0.0001</b>	0.356	<b>0.972</b>	145.32	<b>65.72</b>	

The bold indicates the best performance.



$$\text{TPR} = \frac{\text{Number of correctly estimated edges}}{\text{Number of true edges}},$$

$$\text{FPR} = \frac{\text{Number of incorrectly estimated edges}}{\text{Number of true gaps}}, \text{ and}$$

$$\text{TDR} = \frac{\text{Number of correctly estimated edges}}{\text{Number of estimated edges}},$$

where gaps indicate the pairs of vertex that are not linked.

SHD is the number of edge insertions, deletions, and flips to transfer the estimated DAG into the true DAG [7]. A large SHD indicates a poor fit, whereas a small SHD indicates a good fit.

To simulate the case that the Gaussian assumption is not true, we generated multivariate data with dependency structure by a given DAG with nodes corresponding to random variables and added standard Cauchy distribution in partial samples using the *rmvDAG* function in the *pcalg* package. Figure 6 depicts the normal distribution and the Cauchy distribution. As shown in Figure 6, the standard Cauchy distribution is tail-heavier

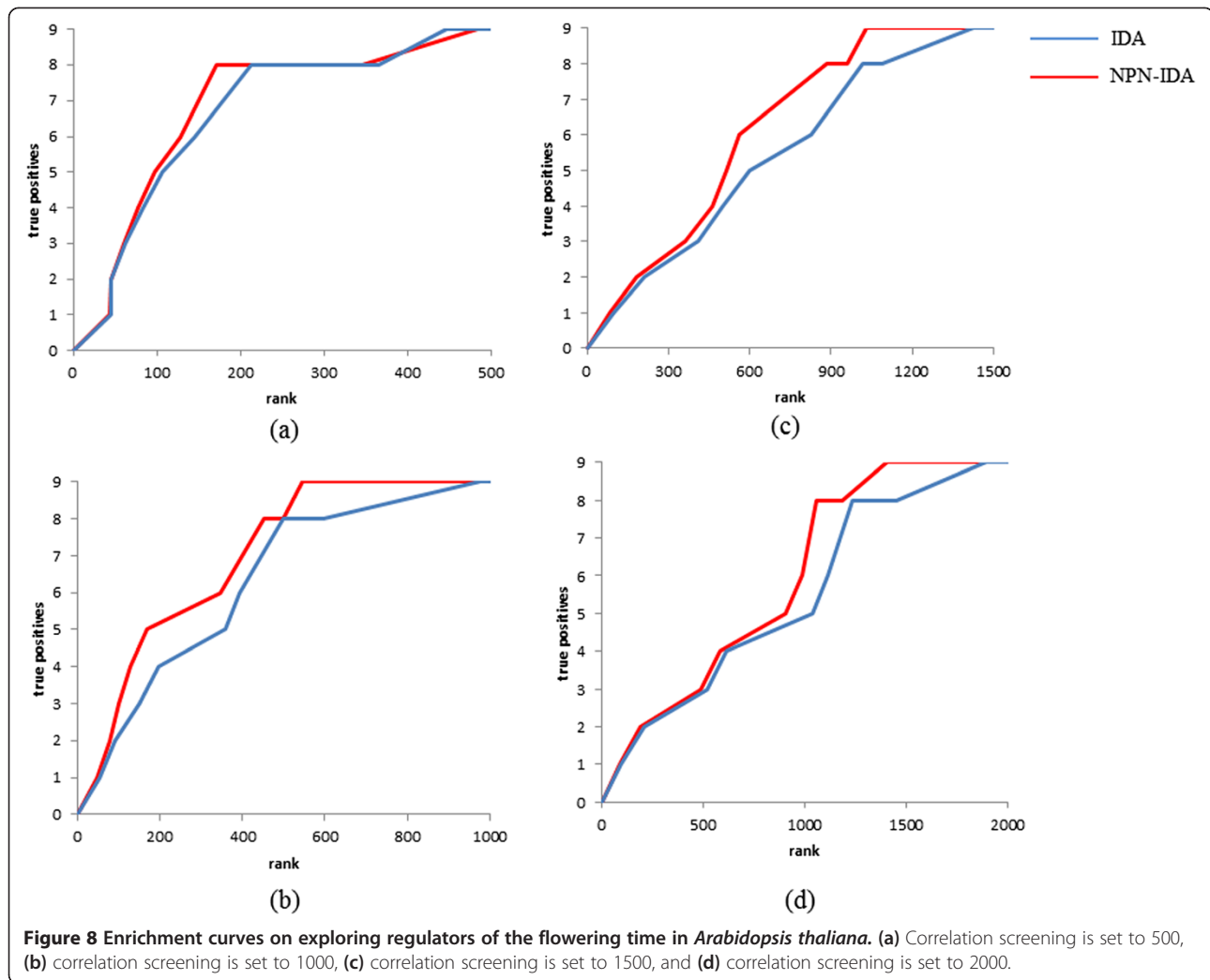
than the standard normal distribution, and is quite different from the standard normal distribution.

The second purpose was to evaluate the performance of NPN-IDA when applied to predicting regulators of the flowering time as phenotype of interest using an *A. thaliana* microarray data set and regulators that control the browning of white adipocytes in mice using mouse microarray data. These two data sets are entirely different in terms of species and target variables. Therefore, if we obtain the consistent results through performance comparison of the methods, the consequence will be solid. We implemented the R language and employed the *pcalg* package for the PC-algorithm and IDA [11].

#### Performance evaluation of the NPN-PC algorithm

We evaluated the performance under all combinations of settings listed in Table 1. The mixing rate indicates the percentage of samples whose error distribution was drawn from the standard Cauchy distribution. The higher the mixing rate, the less accurate is the Gaussianity assumption. To evaluate this hypothesis thoroughly, we repeated this experiment 50 times and averaged the values of TPR, FPR,





TDR, and SHD. To explain the essence of these experiments, we show the representative results of the settings ( $\alpha = 10^{-4}$ ,  $cp = 0.005$ ) in Table 2. All experimental results are shown in Additional file 1. To estimate the causal effects, estimated DAGs inferred by NPN-PC were employed. Because we considered the situation that the Gaussianity assumption is violated, we determined the significance level  $\alpha$  on the basis of the average SHD when the mixing rate was 1. Average SHDs under the different probabilities of

connecting one node to another node are shown in Figure 7. Because the average SHD was very small, we employed estimated DAGs when the significance level  $\alpha$  was set to  $10^{-4}$  to further estimate the causal effects.

#### Performance evaluation of the NPN-IDA algorithm

##### Exploring regulators of the flowering time in *A. thaliana*

We employed the microarray data set of *A. thaliana* and the flowering time used in a previous study [7]. The data set consisted of 21326 probes and 47 samples. We regarded the nine known regulators of the flowering time (*DWF4*, *FLC*, *FRI*, *RPA2B*, *SOC1*, *PDH-E1* BETA,

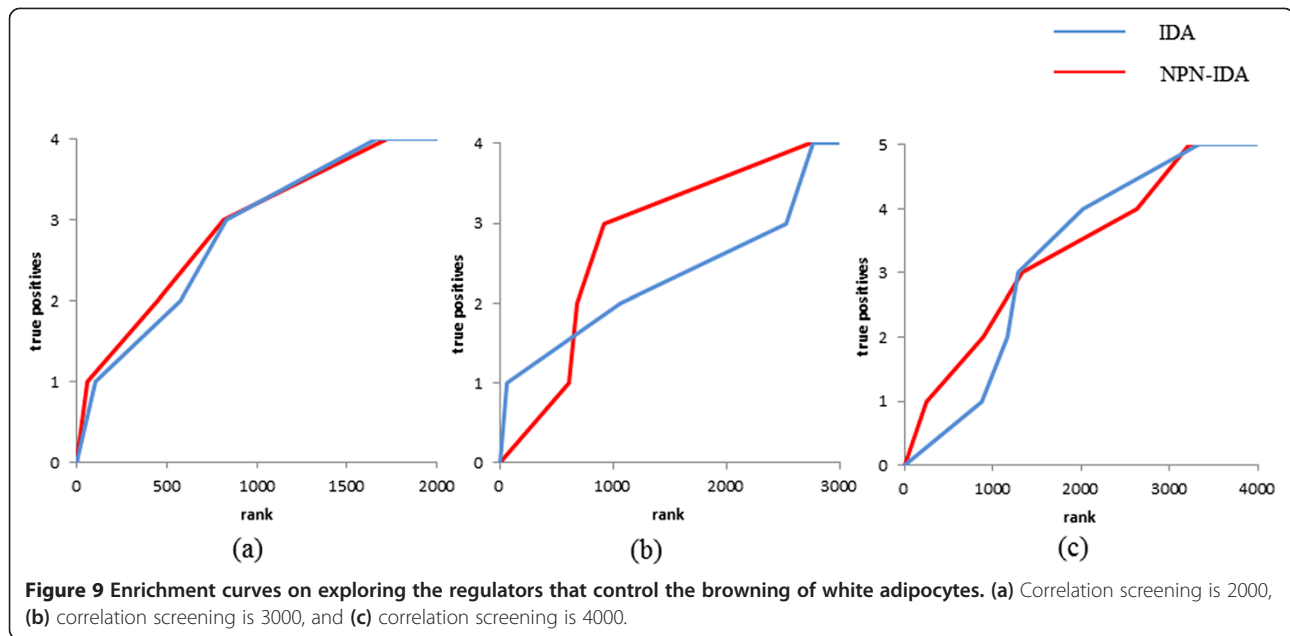
**Table 3** AUC values under the different conditions of correlation screening on exploring regulators of the flowering time in *Arabidopsis thaliana*

Correlation screening	IDA	NPN-IDA
500	0.737	<b>0.755</b>
1000	0.696	<b>0.776</b>
1500	0.605	<b>0.680</b>
2000	0.608	<b>0.662</b>

The bold indicates the best performance.

**Table 4** Known regulators of the differentiation of WAT to BAT

Correlation screening	Gene symbol
2000	<i>Ppargc1a</i> , <i>Tbx15</i> , <i>Tfam</i> , <i>Bmp7</i>
3000	<i>Ppargc1a</i> , <i>Tbx15</i> , <i>Tfam</i> , <i>Bmp7</i>
4000	<i>Foxc2</i> , <i>Ppargc1a</i> , <i>Tbx15</i> , <i>Tfam</i> , <i>Bmp7</i>



*ATGH9B5*, *LRR*, and *OTLD1*) as true-positive genes. Because IDA requires a significant amount of computation time, we selected genes for further estimation that had higher absolute correlation coefficients against flowering time until a predefined number was reached (correlation screening). In this study, the numbers of remaining genes obtained by correlation screening were set to 500, 1000, 1500, and 2000. According to the description in the previous section, we set the significance level  $\alpha$  to  $10^{-4}$  for estimating DAGs. Figure 8 shows the enrichment curves under the different conditions of correlation screening. Vertical axes represent the average numbers of true positives and horizontal axes indicate the ranks of causal effects. To quantitatively compare the performances of IDA and NPN-IDA, we also summarized the area under the enrichment curves (AUCs) under the different conditions of correlation screening in Table 3.

#### Exploring the regulators that control the browning of white adipocytes in mice

We employed the mouse microarray data set of white adipose tissue (WAT) obtained from a previous study [12]. The data set consisted of 43681 probes and 349 WAT samples. According to a previous review [13], we regarded *Ucp1* as marker of brown adipose tissue (BAT). We selected genes that had higher absolute correlation coefficients against *Ucp1* until a predefined number was reached. The numbers of remaining genes were set to 2000, 3000, and 4000. Table 4 shows the known regulators of the differentiation of WAT to BAT for the different conditions of correlation screening. Note that there are no true positive genes when the number of remaining genes obtained by correlation screening is below 1000.

Because there are two probes for *Tbx15*, i.e., *merck-NM\_009323\_at* and *merck-NM\_01154\_a\_at*, we regarded the probe that had the larger causal effect as *Tbx15*. Figure 9 shows the enrichment curves under different conditions of correlation screening. Vertical axes represent the average numbers of true positives and horizontal axes indicate the ranks of causal effects. The AUCs under the different conditions of correlation screening are summarized in Table 5.

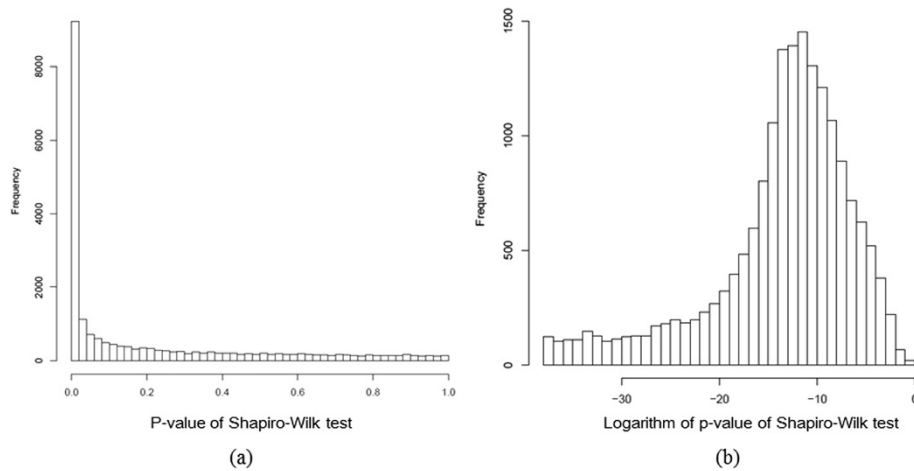
#### Discussion

From Table 2, it appears that the NPN-PC algorithm consistently outperforms the PC-algorithm with regard to all performance metrics, TPR, FPR, TDR, and SHD. Furthermore, as the mixing rate increases, the performance values of the PC-algorithm decrease. This result clearly shows that the PC-algorithm does not work when the Gaussianity assumption is not true. In contrast, the NPN-PC algorithm works well when mixing rate is high. In other words, the NPN-PC algorithm does not require the Gaussianity assumption of data distribution. In terms of FPR, it appears that the FPR of the NPN-PC algorithm is strictly suppressed under all set values. In the NPN-PC algorithm, all performance metrics improved when the number of samples was 100

**Table 5** AUC values under the different conditions of correlation screening on exploring the regulators that control the browning of white adipocytes in mice

Correlation screening	Known regulators	IDA	NPN-IDA
2000	4	0.706	<b>0.726</b>
3000	4	0.580	<b>0.701</b>
4000	5	0.648	<b>0.665</b>

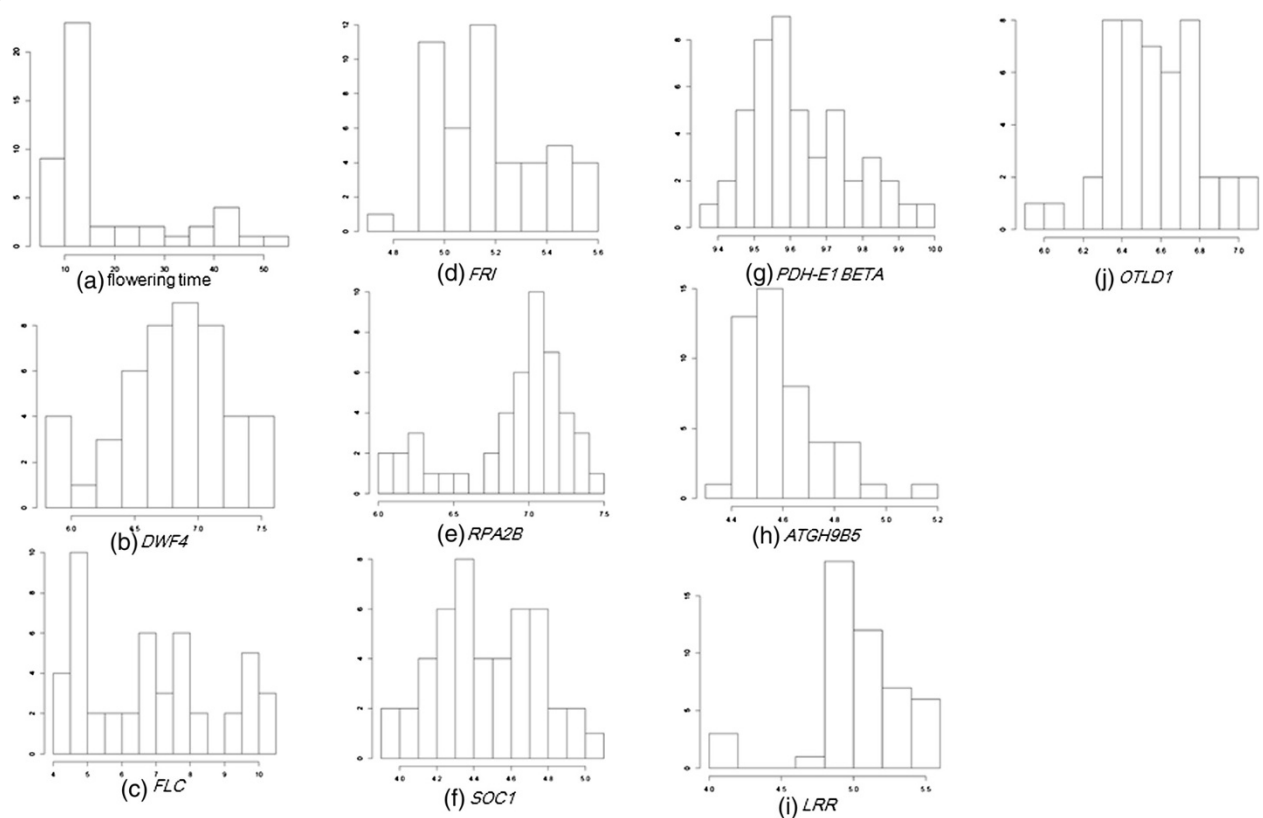
The bold indicates the best performance.



**Figure 10 Histogram of p-values of the Shapiro-Wilk test.** (a) *Arabidopsis thaliana* microarray data and (b) mouse WAT microarray data (logarithmic scale).

compared to when the number of samples was 50. From these observations, it can be concluded that the NPN-PC algorithm is robust and does not depend on data distribution, unlike the PC-algorithm. This characteristic is very attractive from the practical point of view.

From Table 3 and Figure 8, NPN-IDA consistently outperformed IDA on exploring regulators of the flowering time in *A. thaliana*. In particular, when the correlation screenings were set to 1000 and 1500, the difference in the AUCs between NPN-IDA and IDA increased. When the correlation



**Figure 11 Distributions of flowering time and expression levels of the nine known regulators of the flowering time.** Each panel is a histogram of the flowering time and each gene's expression.

screenings were set to 500, both NPN-IDA and IDA worked well. This result indicates that the known regulators were sufficiently recovered against the flowering time within genes that have high absolute correlation coefficients. Therefore, although we do not know whether the unknown regulators have high absolute correlation coefficients against the flowering time, it would be a good strategy from a practical perspective to explore novel regulators from genes that have high absolute correlation coefficients against the flowering time.

From Table 5 and Figure 9, NPN-IDA consistently outperformed IDA on exploring regulators that control the browning of white adipocytes in mice. In particular, when the correlation screenings were set to 3000 and 4000, the difference in the AUCs between NPN-IDA and IDA increased. These results suggest that NPN-IDA successfully enriches known regulators at the top ranks when the number of available genes increases. Consequently, our two experiments clearly demonstrated that NPN-IDA performs better than IDA.

To discuss whether the Gaussian assumption is true or not in the data sets used in this study, we applied the Shapiro-Wilk test to microarray data and a target variable of interest, which tests the null hypothesis that a sample came from a normally distributed population [14]. We show a histogram of the p-values of the Shapiro-Wilk test for target phenotypes of interest (flowering time in *A. thaliana* and gene expression of *Ucp1* in mice) and gene expression levels in Figure 10. We also show individual histograms of phenotypes of interest and expression levels of the known regulators in Figure 11; the respective p-values of the Shapiro-Wilk test are summarized in Table 6. From Figure 10, it appears that the p-values of most genes were <0.05 for both *A. thaliana* and mouse WAT microarray data. In other words, the normality assumption was violated in most genes. These results justify the use of NPN-IDA rather than IDA, because the latter method requires a normal distribution. With regard to the *A. thaliana* data, it appears that the p-values for flowering time, *FLC*, *FRI*, *RPA2B*, *ATGH9B5*, and *LRR* were <0.05 (Table 6); as shown in Figure 11, their distributions were skewed. With regard to the mouse WAT data, the p-values of all genes were very small (Table 6). As shown in Figure 12, their distributions were also skewed. These results strongly suggest that NPN-IDA, which does not require the Gaussian distribution, works better than IDA.

Although the method presented here performs significantly better than IDA, one might consider the difference in the performance as too small. However, within the known regulators of flowering time in *A. thaliana*, four genes (*PDH-E1 BETA*, *ATGH9B5*, *LRR*, and *OTLD1*) were experimentally validated using the IDA-based method [6]. Therefore, one should take into account that some of the known regulators were obtained on the basis of the Gaussian

**Table 6 p-values of the Shapiro-Wilk test of target variables and known regulators**

Variable	p-value
Flowering time	1.09e-7
<i>DWF4</i>	3.62e-1
<i>FLC</i>	1.81e-3
<i>FRI</i>	2.81e-1
<i>RPA2B</i>	1.24e-4
<i>SOC1</i>	6.86e-1
<i>PDH-E1 BETA</i>	7.95e-2
<i>ATGH9B5</i>	1.69e-4
<i>LRR</i>	1.21e-4
<i>OTLD1</i>	6.84e-1
<i>Ucp1</i>	1.50e-21
<i>Foxc2</i>	6.01e-22
<i>Ppargc1a</i>	8.59e-21
<i>Tbx15 (merck-NM_00923_at)</i>	1.75e-21
<i>Tbx15 (merck-NM_01154_a_at)</i>	6.31e-20
<i>Tfam</i>	5.12e-08
<i>Bmp7</i>	3.09e-13

assumption. Thus, it is possible that the improvement achieved with NPN-IDA is greater than the experimental results shown in this study.

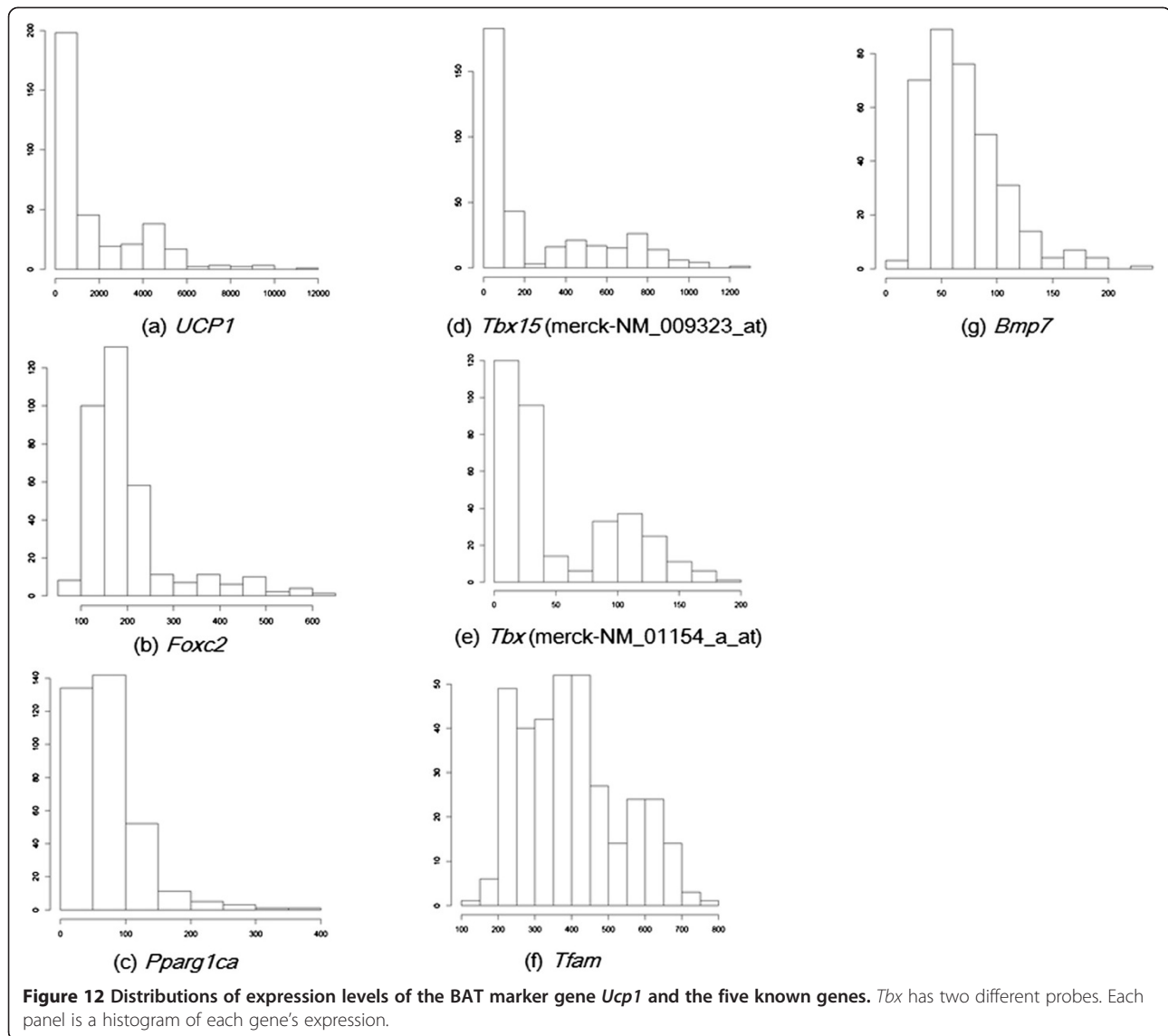
In summary, the two main results of our experimental study are: (1) the NPN-PC algorithm works better than the PC-algorithm in estimating DAGs on synthetic data, and (2) NPN-IDA performs better than does IDA in predicting regulators of the flowering time in *A. thaliana* and regulators of the differentiation of WAT to BAT in mice on the basis of microarray data. From these two results, we conclude that the performance to estimate DAGs contributes to the accurate prediction of causal effects.

From a practical point of view, because regulators that control the browning of white adipocytes have not been identified only from microarray data of WAT so far, it might be worthwhile to demonstrate this possibility for the first time using our novel method.

For further performance enhancement, we consider that NPN-IDA could be embedded into CStAR (causal stability ranking) [6] in the future. CStAR uses IDA with stability selection [6,15] and the performance was greatly improved compared to plain IDA. By simply replacing IDA with NPN-IDA in the estimation process for causal effects, it would be easy to combine NPN-IDA with CStAR.

## Conclusions

We presented NPN-IDA, which uses nonparametric partial correlations, to test conditional independencies in the PC-algorithm for intervention-calculus. To reveal



the contribution of estimating DAGs, we evaluated the NPN-PC algorithm to estimate DAGs with the *non-paranormal* method. The two main results of our experimental study are: (1) the NPN-PC algorithm works better than the PC-algorithm in estimating DAGs on synthetic data, and (2) NPN-IDA performs better than IDA does in predicting regulators of the flowering time in *A. thaliana* and regulators of the differentiation of WAT to BAT in mice. Considering these two results, we concluded that the performance to estimate DAGs contributes to the accurate prediction of causal effects.

Thus, the simplest alternative procedure we used for our proposed method enables us to design efficient intervention experiments and can be applied to a wide range of research purposes, including drug discovery and medicine, because of its generality.

#### Availability of supporting data

The microarray data sets used in this study are deposited in Additional file 2.

#### Additional files

**Additional file 1:** Experimental results of all combined parameter settings for performance comparison between the NPN-PC algorithm and the PC-algorithm.

**Additional file 2:** The mouse microarray data set of microarray data set of white adipose tissue (WAT) and the microarray data set of *A. thaliana* and the flowering time.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

RT proposed the method, implemented, conduct experiments and wrote the manuscript. CS edited the tables. SF supervised the research project. All authors read and approved the final manuscript.

#### Acknowledgements

We are grateful to Tatsumi Yamazaki, the president of Forerunner Pharma Research, for his support and encouragement throughout the study. We also thank Takumi Nagata and Kenji Yoshida for their valuable discussions.

Received: 12 March 2014 Accepted: 24 June 2014

Published: 30 June 2014

#### References

1. Ahfeldt T, Schinzel RT, Lee YK, Hendrickson D, Kaplan A, Lum DH, Camahort R, Xia F, Shay J, Rhee EP, Clish CB, Deo RC, Shen T, Lau FH, Cowley A, Mowrer G, Al-Siddiqi H, Nahrendorf M, Musunuru K, Gerszten RE, Rinn JL, Cowan CA: **Programming human pluripotent stem cells into white and brown adipocytes.** *Nat Cell Biol* 2012, **14**:209–219.
2. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S: **Induction of pluripotent stem cells from adult fibroblasts by defined factors.** *Cell* 2007, **13**:861–872.
3. Ring KL, Tong LM, Balestra ME, Javier R, Andrew-Zwilling Y, Li G, Walker D, Zhang WR, Kreitzer AC, Huang Y: **Direct reprogramming of mouse and human fibroblasts into multipotent neural stem cells with a single factor.** *Cell Stem Cell* 2012, **11**:100–2109.
4. Maathuis MH, Kalisch MK, Buhlmann P: **Estimating high-dimensional intervention effects from observational data.** *Ann Stat* 2009, **37**:3133–3164.
5. Maathuis MH, Colombo D, Kalisch MK, Buhlmann P: **Predicting causal effects in large-scale systems from observational data.** *Nat Methods* 2010, **7**:247–248.
6. Stekhoven DJ, Moraes I, Sveinbjornsson G, Hennig L, Maathuis MH, Buhlmann P: **Causal stability ranking.** *Bioinformatics* 2012, **28**:2819–2823.
7. Kalisch M, Buhlmann P: **Estimating high-dimensional directed acyclic graphs with the PC-algorithm.** *J Machine Learning Res* 2007, **8**:613–636.
8. Liu H, Laffery J, Wasserman L: **The nonparanormal: semiparametric estimation of high dimensional undirected graphs.** *J Machine Learning Res* 2009, **10**:2295–2328.
9. Liu H, Han F, Yuan M, Laffery J, Wasserman L: **The Nonparanormal SKEPTIC.** In *Proceedings of the 29th International Conference of Machine Learning*. Edinburg, Scotland, UK: 2012:1415–1422.
10. Kruskal WH: **Ordinal measures of association.** *J Am Stat Assoc* 1958, **53**:814–861.
11. Kalisch M, Maechler M, Colombo D, Maathuis MH, Buhlmann P: **Causal inference using graphical models with R package pcalg.** *J Stat Software* 2012, **47**:1–26.
12. MuiSE ES, Souza S, Chi A, Tan Y, Zhao X, Liu F, Dallas-Yang Q, Wu M, Sarr T, Zhu L, Guo H, Li Z, Li W, Hu W, Jiang G, Paweletz CP, Hendrickson RC, Thompson JR, Mu J, Berger JP, Mehmet H: **Downstream signaling pathways in mouse adipose tissues following acute in vivo administration of fibroblast growth factor 21.** *PLoS One* 2013, **8**:e73011.
13. Lo KA, Sun L: **Turning WAT into BAT: a review on regulators controlling the browning of white adipocytes.** *Biosci Rep* 2013, **33**:711–719.
14. Shapiro SS, Wilk MB: **An analysis of variance test for normality (complete samples).** *Biometrika* 1965, **52**:591–611.
15. Meinshausen N, Buhlmann P: **Stability selection.** *J Roy Stat Soc B* 2010, **72**:417–473.

doi:10.1186/1471-2105-15-228

Cite this article as: Teramoto *et al.*: Estimating causal effects with a non-paranormal method for the design of efficient intervention experiments. *BMC Bioinformatics* 2014 **15**:228.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

