**Nimisha Chaturvedi[1,2] / Renée X. de Menezes[1,2] / Jelle J. Goeman[3] / Wessel van Wieringen[1,4]**

# A test for detecting differential indirect trans effects between two groups of samples

[1] Afdeling Epidemiologie en Biostatistiek, Amsterdam Public Health Research Institute, Medische Faculteit (F-vleugel), VU Medisch Centrum, 1007 MB Amsterdam, The Netherlands, E-mail: nimisha.chaturvedi@epfl.ch
[2] Netherlands Bioinformatics Center, 260 NBIC, 6500 HB Nijmegen, The Netherlands, E-mail: nimisha.chaturvedi@epfl.ch
[3] Department of Biomedical Data Sciences, Room Number S5-P, LUMC Main Building, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands
[4] Department of Mathematics, Amsterdam Public Health Research Institute, Faculty of Sciences, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

**Abstract:**
Integrative analysis of copy number and gene expression data can help in understanding the cis and trans effect of copy number aberrations on transcription levels of genes involved in a pathway. To analyse how these copy number mediated gene-gene interactions differ between groups of samples we propose a new method, named dNET. Our method uses ridge regression to model the network topology involving one gene's expression level, its gene dosage and the expression levels of other genes in the network. The interaction parameters are estimated by fitting the model per gene for all samples together. However, instead of testing for differential network topology per gene, dNET tests for an overall difference in estimated parameters between two groups of samples and produces a single $p$-value. With the help of several simulation studies, we show that dNET can detect differential network nodes with high accuracy and low rate of false positives even in the presence of differential cis effects. We also apply dNET to publicly available TCGA cancer datasets and identify pathways where copy number mediated gene-gene interactions differ between samples with cancer stage lower than stage 3 and samples with cancer stage 3 or above.

**Keywords:** group testing, high dimensional data, multivariate analysis, network analysis
**DOI:** 10.1515/sagmb-2017-0058

## 1 Introduction

The accumulation of genetic aberrations is known to drive the progression of diseases such as cancer by affecting the regulating mechanism of the cell. For example, copy number changes for gene $j$ not only affects its own transcript levels (cis-effect) but also of other genes working together with it as a unit in a pathway (indirect trans effect). To understand these complex association patterns within a pathway, it is important to integrate information from different molecular profiles at geneset level. With a method for inferring the copy number mediated network topology, the next interesting question will be whether the topology differs between groups of samples?

Although some advanced and very efficient methods are available to study these associations patterns between different molecular levels for the same set of samples (Chaturvedi et al., 2014; Chaturvedi, Menezes & Goeman, 2017; Flutre et al., 2013; Horlings et al., 2010; Menezes et al., 2009; van Iterson et al., 2013; van Wieringen, Berkhof & van de Wiel, 2010), they often suffer from one or more of the three major drawbacks. First, they cannot test whether the association patterns are phenotype specific. Second, they integrate at the level of genes leading to problems due to large search space such as low power due to multiple testing burden (Kendziorski et al., 2006; Listgarten, Kadie & Heckerman, 2010). Third, they are not designed to model the network topology.

Recently, van Wieringen & van de Wiel, 2014 proposed an integration method for analysing the indirect trans-effects (influence of abnormal gene dosage for, say, gene $i$ over the transcript levels of gene $j$ via the mRNA and protein levels of gene $i$) within a gene regulatory pathway. In this work, they demonstrate that integrating DNA copy number data with gene transcript levels benefits the discovery of gene-gene interactions. One naive way to compare the pathway topology between two groups of samples would be to use this method for estimating the model parameters, that define the network nodes, separately for two groups of samples and then compare them. However, comparing the association effects estimated for two group of samples separately can

**Nimisha Chaturvedi** is the corresponding author.

suffer from various issues such as loss of power due to analysing small subset of samples one at a time and lack of interpretability due to combining and interpreting results from different analysis.

To tackle the problems of separate group estimation Chaturvedi et al. (2014) proposed a novel method, named dSIM. This method tests for differences in copy number led gene expression regulation between two groups of samples while using all data together, thus eliminating the drawbacks of separate analysis. It also reduces the number of tests by modelling association between every feature in one dataset and sets of covariates in the other. However, this method is not designed for analysing indirect trans-effects in a pathway. Moreover, since dSIM can only handle a univariate response, it can still suffer from low power due to multiple testing burden.

In this paper, we propose a novel multivariate test, named dNET, for detecting differential indirect trans effects between two groups of samples. The method uses a rate equation based model (van Wieringen & van de Wiel, 2014) with ridge penalization (Hoerl & Kennard, 1970) for estimating the parameters while fitting a single model on all samples together. These estimates are used to obtain a single multivariate test statistic between two sets of variables and the final $p$-value is obtained by using permutation testing. By considering a multivariate response, dNET reduces the total number of tests to one, making the analysis more powerful and less time consuming. It is, therefore, useful for detecting differential effects on the pathway level before exploring individual nodes. Following sections describe the derivation of the test and various steps involved in calculating the $p$-value. To measure the sensitivity and specificity of dNET under various conditions, we performed several simulation studies and also analysed publicly available TCGA (The Cancer Genome Atlas) cancer datasets.

## 2 Methods

### 2.1 Motivation

Available are gene expression and copy number data on $n$ samples and $p$ genes involved in a pathway, denoted as $\mathbf{Y}_{n\times p}$ and $\mathbf{X}_{n\times p}$, respectively. The samples can be divided into two groups based on some binary phenotype, i.e. a sample is either (say) $a$ or $b$. Each phenotype being present with $n^{(a)}$ and $n^{(b)}$ samples and $n = n^{(a)} + n^{(b)}$. Every sample of either phenotype $a$ and $b$ the indirect *trans*-effect per gene is modelled using the regression model

$$\begin{cases} y_j^{(a)} &= Y_{-j}^{(a)}(-\theta_j^{(a)}) + x_j^{(a)}\beta_j + \varepsilon_j^{(a)} \quad j = 1,\dots,p, \\ y_j^{(b)} &= Y_{-j}^{(b)}(-\theta_j^{(b)}) + x_j^{(b)}\beta_j + \varepsilon_j^{(b)} \quad j = 1,\dots,p, \end{cases} \tag{1}$$

where $y_j^{(a)}$ and $y_j^{(b)}$ are $(n^{(a)}, n^{(b)}) \times 1$ vectors of expression values for the $j$th gene, $Y_{-j}^{(a)}$ and $Y_{-j}^{(b)}$ are $(n^{(a)}, n^{(b)}) \times (p-1)$ matrices of expression values for all genes, except the $j$th one and $x_j^{(a)}$ and $x_j^{(b)}$ are $(n^{(a)}, n^{(b)}) \times 1$ vectors of copy number values for the $j$th gene (van Wieringen & van de Wiel, 2014; Jornsten et al., 2011). The $(p-1) \times 1$ vectors of coefficients, $-\theta_j^{(a)}$ and $-\theta_j^{(b)}$, represents the indirect *trans*-effects and $\beta_j$, of dimensions $1 \times 1$ represents the association coefficient between gene expression and copy number for gene $j$. We assume that $(\varepsilon_1^{(a)},\dots,\varepsilon_p^{(a)})$ and $(\varepsilon_1^{(b)},\dots,\varepsilon_p^{(b)})$ are independently distributed with $\varepsilon_j^{(a)} \sim \mathcal{N}(0, \sigma_j^2 I_{n_a})$ and $\varepsilon_j^{(b)} \sim \mathcal{N}(0, \sigma_j^2 I_{n_b})$.

The main objective of this work is to test for an overall difference in the indirect *trans*-effects within a pathway, when the samples are grouped according to some binary grouping variable (for example: metastasized or not). A naive approach to detect differences between $-\theta_j^{(a)}$ and $-\theta_j^{(b)}$ would be to analyse groups of samples separately by estimating $-\theta_j$ per gene, test for indirect *trans*-effects and then comparing the results between both groups. However, testing for differential effects per gene for each group separately may result into low power due to small sample sizes as well as multiple testing correction. Moreover, analysing the subsets separately will require an additional step of meta-analysing the two distinct sets of $p$-values.

Our new approach dNET overcomes these challenges by estimating the parameters $-\theta_j$ for all samples together and then testing for differences between two sets of variables, producing a single $p$-value.

### 2.2 Modelling the difference between the phenotypic groups

Here we describe the derivation of the test that can detect differences in overall association between two groups of samples. To this end our model is rewritten in matrix notation. Define for phenotype $a$:

$$\mathbf{Y}^{(a)} = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n_a 1} & \cdots & y_{n_a p} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{1^*} \\ \vdots \\ \mathbf{Y}_{n_a *} \end{pmatrix},$$

$$\mathbf{X}^{(a)} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n_a 1} & \cdots & x_{n_a p} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{1^*} \\ \vdots \\ \mathbf{X}_{n_a *} \end{pmatrix},$$

and $\boldsymbol{\beta} = \text{diag}(\beta^1, \ldots, \beta^p)$. Similar matrices are defined for phenotype $b$. Following the assumptions stated in van Wieringen and van de Wiel (2014) the models for both phenotypes can then be written as

$$\begin{cases} \boldsymbol{\Theta}^{(a)}\mathbf{Y}^{(a)^\top} = \boldsymbol{\beta}\mathbf{X}^{(a)^\top} + \boldsymbol{\varepsilon}^{(a)^\top} \\ \boldsymbol{\Theta}^{(b)}\mathbf{Y}^{(b)^\top} = \boldsymbol{\beta}\mathbf{X}^{(b)^\top} + \boldsymbol{\varepsilon}^{(b)^\top} \end{cases} \tag{2}$$

where $\boldsymbol{\varepsilon}^{(a)^\top} = (\varepsilon_1^{(a)}, \ldots, \varepsilon_p^{(a)})^\top$ and $\boldsymbol{\varepsilon}^{(b)^\top} = (\varepsilon_1^{(b)}, \ldots, \varepsilon_p^{(b)})^\top$. The indirect *trans*-effects for $g_a$ and $g_b$ are represented by $\boldsymbol{\Theta}^{(a)}$ and $\boldsymbol{\Theta}^{(b)}$, respectively and $\boldsymbol{\Sigma}$ represents the dispersion matrix with variance per gene in the diagonal. Assume that the common indirect *trans*-effects and the differential indirect trans effects between the two phenotypes $a$ and $b$ are represented by $p \times p$ matrices $\boldsymbol{\Theta}$, with $\text{diag}(\boldsymbol{\Theta}) = \mathbf{1}_p$, and $\boldsymbol{\delta}$, with $\text{diag}(\boldsymbol{\Theta}) = \mathbf{0}_p$, respectively. To define these matrices in terms of $\boldsymbol{\Theta}^{(a)}$ and $\boldsymbol{\Theta}^{(b)}$ we can write, $\frac{1}{2}(\boldsymbol{\Theta}^{(a)} + \boldsymbol{\Theta}^{(b)}) = \boldsymbol{\Theta}$ and $\frac{1}{2}(\boldsymbol{\Theta}^{(a)} - \boldsymbol{\Theta}^{(b)}) = \boldsymbol{\delta}$ which in turn gives,

$$\boldsymbol{\Theta}^{(a)} = (\boldsymbol{\Theta} + \boldsymbol{\delta}) \quad \text{and} \quad \boldsymbol{\Theta}^{(b)} = (\boldsymbol{\Theta} - \boldsymbol{\delta}) \tag{3}$$

Replacing $\boldsymbol{\Theta}^{(a)}$ and $\boldsymbol{\Theta}^{(b)}$ with $(\boldsymbol{\Theta} + \boldsymbol{\delta})$ and $(\boldsymbol{\Theta} - \boldsymbol{\delta})$ in 4, results in the equivalent,

$$\begin{cases} (\boldsymbol{\Theta} + \boldsymbol{\delta})\mathbf{Y}^{(a)^\top} = \boldsymbol{\beta}\mathbf{X}^{(a)^\top} + \boldsymbol{\varepsilon}^{(a)^\top} \\ (\boldsymbol{\Theta} - \boldsymbol{\delta})\mathbf{Y}^{(b)^\top} = \boldsymbol{\beta}\mathbf{X}^{(b)^\top} + \boldsymbol{\varepsilon}^{(b)^\top} \end{cases} \tag{4}$$

Here columns of $\boldsymbol{\varepsilon}^{(a)}$ and $\boldsymbol{\varepsilon}^{(b)}$ are i.i.d. $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

## 2.3 Testing for a difference

We are interested in testing the null hypothesis $H_0 : \boldsymbol{\delta} = \mathbf{0}_{pp}$. Since the alternative is high-dimensional, we use the theory of Goeman, van De Geer, and van Houwelingen (2006) who advocate a score test defined as

$$S = \mathbf{u}^\top \mathbf{u} - \text{trace}(\mathbf{V}), \tag{5}$$

where $\mathbf{u}$ is the score vector and $\mathbf{V}$ is the Fisher information matrix of $\boldsymbol{\delta}$ under $H_0$. A test based on (5) has the property that it is locally most powerful on average in a neighbourhood of the null hypothesis. When calculating $\mathbf{u}$ and $\mathbf{V}$ we may ignore all terms in the loglikelihood that do not involve both $\boldsymbol{\delta}$ and $\mathbf{Y}$, since this leads to an equivalent test statistic.

The model implies that, e.g.

$$\mathbf{Y}^{(a)}|\mathbf{X}^{(a)} \sim \mathcal{N}\{(\boldsymbol{\Theta} - \boldsymbol{\delta})^{-1}\boldsymbol{\beta}\mathbf{X}^{(a)}, (\boldsymbol{\Theta} - \boldsymbol{\delta})^{-1}\boldsymbol{\Sigma}[(\boldsymbol{\Theta} - \boldsymbol{\delta})^{-1}]^\top\}, \tag{6}$$

and similarly for $\mathbf{Y}^{(b)} \mid \mathbf{X}^{(b)}$. The log-likelihood, after some linear algebra and leaving out terms not involving the parameters, is:

$$\log L_0 \propto n_a \log|\boldsymbol{\Theta} + \boldsymbol{\delta}| + n_b \log|\boldsymbol{\Theta} - \boldsymbol{\delta}| - \frac{1}{2}\text{trace}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}\mathbf{L}\boldsymbol{\delta}^\top)\} + \text{trace}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}\mathbf{E})\}, \tag{7}$$

where, $\mathbf{L} = \mathbf{Y}^{(a)^\top}\mathbf{Y}^{(a)} + \mathbf{Y}^{(b)^\top}\mathbf{Y}^{(b)}$ and $\mathbf{E} = \mathbf{E}^{(b)} - \mathbf{E}^{(a)}$ with $\mathbf{E}^{(a)} = \mathbf{Y}^{(a)^\top}(\mathbf{Y}^{(a)}\boldsymbol{\Theta}^\top - \mathbf{X}^{(a)}\boldsymbol{\beta})$ and $\mathbf{E}^{(b)} = \mathbf{Y}^{(b)^\top}(\mathbf{Y}^{(b)}\boldsymbol{\Theta}^\top - \mathbf{X}^{(b)}\boldsymbol{\beta})$. Removing terms not involving both $\boldsymbol{\delta}$ and $\mathbf{Y}$, we only need

$$\log L_0 \propto -\frac{1}{2}\text{trace}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}\mathbf{L}\boldsymbol{\delta}^\top)\} + \frac{1}{2}\text{trace}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}\mathbf{E})\}.$$

To calculate the score we use the trace property of matrix derivatives:

$$\partial\text{trace}(\mathbf{AXBX}^\top\mathbf{C}) = \mathbf{B}^\top\mathbf{X}^\top\mathbf{A}^\top\mathbf{C}^\top\partial\mathbf{X} + \mathbf{BX}^\top\mathbf{CA}\partial\mathbf{X}.$$

Using this to calculate first derivatives w.r.t $\boldsymbol{\delta}$ we can take

$$\frac{\partial}{\partial\boldsymbol{\delta}} lnL_0 \propto \boldsymbol{\Sigma}^{-1}\mathbf{E}^\top - \mathbf{L}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}, \tag{8}$$

so that $\mathbf{u} = \text{vec}(\boldsymbol{\Sigma}^{-1}\mathbf{E}^\top)$.

For obtaining the second derivative we note that the first term does not involve $\delta$. To obtain the derivative of the second term we define $\mathbf{Q} = \boldsymbol{\Sigma}^{-1} = diag(\sigma_1^{-2}, \ldots, \sigma_p^{-2})$. Then, the $(j_1, j_2)$th element of $\mathbf{Q}\boldsymbol{\delta}\mathbf{L}^\top$ is $q_{j_1}^\top\boldsymbol{\delta}l_{j_2}$, where $q_{j_1}^\top = (0, \ldots, \sigma_{j_1}^{-2}, 0, \ldots, 0)$ is the $j_1$th row of $\mathbf{Q}$ and $l_{j_2}$ is the $j_2$th column of the matrix $\mathbf{L}$. The differentiation of $q_{j_1}^\top\boldsymbol{\delta}l_{j_2}$ w.r.t. $\boldsymbol{\delta}$ can be written as $\frac{\partial}{\partial\boldsymbol{\delta}}q_{j_1}^\top\boldsymbol{\delta}l_{j_2} = \frac{\partial}{\partial\boldsymbol{\delta}}\text{trace}(q_{j_1}^\top\boldsymbol{\delta}l_{j_2}) = q_{j_1}l_{j_2}^\top$, which is a $p \times p$ matrix. Using this the differentiation of $\mathbf{Q}\boldsymbol{\delta}\mathbf{L}^\top$ w.r.t $\boldsymbol{\delta}$ can be given as

$$\mathbf{V} = -\frac{\partial}{\partial\boldsymbol{\delta}}\mathbf{Q}\boldsymbol{\delta}\mathbf{L}^\top = -\begin{pmatrix} q_1l_1^\top & \cdots & q_1l_p^\top \\ \vdots & \ddots & \vdots \\ q_pl_1^\top & \cdots & q_pl_p^\top \end{pmatrix}.$$

The matrix $\mathbf{V}$ is a $p^2 \times p^2$ matrix which is costly to calculate. However, we only need its trace, which is $\text{trace}(\mathbf{V}) = -\sum_{j=1}^p l_{jj}\sigma_j^{-2}$, where $l_{jj}$ is the $j$th diagonal element of $\mathbf{L}$.

## 2.4 Estimation of parameters

To obtain estimates of the matrices $\boldsymbol{\Theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ under the null hypothesis we start with equation-by-equation estimation of the parameters for all samples together. For this, we use ridge regression to regress the expression of gene $i$, on that of all other genes and its own DNA copy number. The ridge regression coefficients for the model with all samples together can be defined as

$$\begin{aligned}\theta_j^{ridge} &= \text{argmin}_{\beta,\theta}(y_j - Y_{-j}(-\theta_j) - x_j\beta_j)^\top(y_j - Y_{-j}(-\theta_j) - x_j\beta_j) \\ &\quad + \lambda_j(-\theta_j)^\top(-\theta_j) \quad j = 1, \ldots, p.\end{aligned} \tag{9}$$

Following van Wieringen and van de Wiel (2014) we penalise only $\theta_j$ and not $\beta_j$. To deal with the problem of over-fitting we optimise the ridge penalty parameter $\lambda$ by leave-one-out cross-validation on a grid of $\lambda$ values, maximising the cross-validated likelihood. Then the optimal $\lambda$ is used to obtain the final parameter estimates. The residual variance yields an estimate of $\sigma_j$. We repeat this for $j = 1, \ldots, p$, each time obtaining a row of $\boldsymbol{\Theta}$ and an element of $\boldsymbol{\beta}$.

## 2.5 Permutation *p*-values

To avoid parametric assumptions about the joint distribution of the test statistic we use permutation testing by permuting the group labels $(a, b)$. For each permutation we permute $\mathbf{X}$ and $\mathbf{Y}$ together. Once we have all the permuted test statistics $\{S_1, \ldots S_Z\}$ and the observed test statistic $S$, the permutation $p$-value is then obtained (Phipson & Smyth, 2010) by

$$\boldsymbol{p}_{perm} = \frac{\sum_{\ell=1}^Z \mathbf{I}(S_\ell \geq S) + 1}{Z + 1}$$

where **I(.)** is the indicator function. Since the estimates of nuisance parameters $\boldsymbol{\Theta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ do not involve the group labels, there is no need to re-estimate them for every permutation.

# 3 Results

## 3.1 Application on simulated datasets

### 3.1.1 Simulation setup

We analysed three main features of dNET using our simulation setup. First, the sensitivity of dNET towards detecting pathways with differential nodes (Simulation study 1). Second, the effect of total sample size and the difference in group size on the accuracy of dNET (Simulation study 2). Third, the effect of differential cis-associations on dNET's specificity (Simulation study 3).

For simplicity, we considered the setting $m = n$ and $p = q$ and pre-assigned the samples into groups $a$ and $b$. As a first step, we generated copy number profiles $\mathbf{X}^{(a)}$, $\mathbf{X}^{(b)}$ for n probes from a normal distribution with mean 0. Out of all samples in group $a$, 50% randomly chosen samples were assigned values to a predefined stretch of probes from a normal distribution with mean $\neq 0$. Similar step was repeated for group $b$ also. This structure simulated a copy number data, where only the selected samples received copy number aberrations in the specified region.

For simulating gene expression values in $\mathbf{Y}^{(a)}$ and $\mathbf{Y}^{(b)}$ under simulation studies 1 and 2, we started with generating the data from a normal distribution with mean = 0. To mimic the indirect trans associations within a gene expression network we estimated the $\mathbf{\Theta}$ matrix, as explained in subsection 2.4, from the TCGA bladder cancer gene expression dataset and then used it for incorporating effects in the simulated datasets. We generated datasets with similar association patterns within a gene expression network, between two groups (scenario 1) as well as datasets with differential association patterns within the network, between two groups of samples (scenario 2). Under scenario 1, we used $\mathbf{\Theta}^{(a)} = \mathbf{\Theta}^{(b)} = \mathbf{\Theta}$ and simulated the gene expression networks for sample groups $a$ and $b$ as $\mathbf{Y}^{(a)} = -\mathbf{\Theta}^{(a)}\mathbf{Y}^{(a)}$ and $\mathbf{Y}^{(b)} = -\mathbf{\Theta}^{(b)}\mathbf{Y}^{(b)}$. For scenario 2, we simulated associations within the gene expression network only for samples in group $a$ with $\mathbf{\Theta}^{(a)} = \mathbf{\Theta}$, creating differential indirect trans associations between groups $a$ and $b$.

To analyse the effect of cis-associations on dNET in Simulation study 3, we simulated gene expression values in $\mathbf{Y}^{(a)}$ and $\mathbf{Y}^{(b)}$ under three different cases. Case 1, when there is no association between $\mathbf{Y}$ and $\mathbf{X}$. Case 2, the associations between $\mathbf{Y}$ and $\mathbf{X}$ differ between groups $a$ and $b$. Case 3, when the associations between $\mathbf{Y}$ and $\mathbf{X}$ are similar between samples in groups $a$ and $b$. Out of all the samples in group $a$ or $b$, only those samples with copy number aberrations could have associations with gene expression data. From the same region where copy number aberrations were assumed, probes were chosen with a given probability for exhibiting associational patterns. For all selected probes gene expression values were generated as,

$$Y_{ij}^{G} = \beta^{G} X_{ij}^{G} + \epsilon, \tag{10}$$

where $j \in S \in G$ with $S$ =a vector containing the indices of the samples with copy number aberrations belonging to $G = a$ or $G = b$. $\beta^{G}$ is a single coefficient value generated from a normal distribution with mean $\neq 0$. This value remains the same for all $i$ and $j$. For adding noise to the data, an error value $\epsilon$ is added generated from a normal distribution with mean =0. For case 2, cis-associations were simulated only for samples in group $a$ with $diag(\boldsymbol{\beta}^{(a)}) = \beta$. For case 3, we kept $diag(\boldsymbol{\beta}^{(a)}) = diag(\boldsymbol{\beta}^{(b)}) = \beta$ to have similar cis-associations between two groups of samples.

### 3.1.2 Simulation study 1

Under this simulation study, we tested the sensitivity and specificity of dNET towards detecting differential effects. To do this we simulated the datasets under scenario 1 and scenario 2. For each scenario, we generated 100 data sets ($\mathbf{X}$ and $\mathbf{Y}$) with 200 samples (100 samples per group) and 77 genes.

It can be seen from Figure 1 that dNET can detect the differential effects with very high sensitivity. With the significance threshold set at 0.01, dNET gives significant $p$-values for all 100 datasets with simulated differential effects. The test also gives 100% specificity, with no false positive $p$-values in the scenario when there are no differential effects.

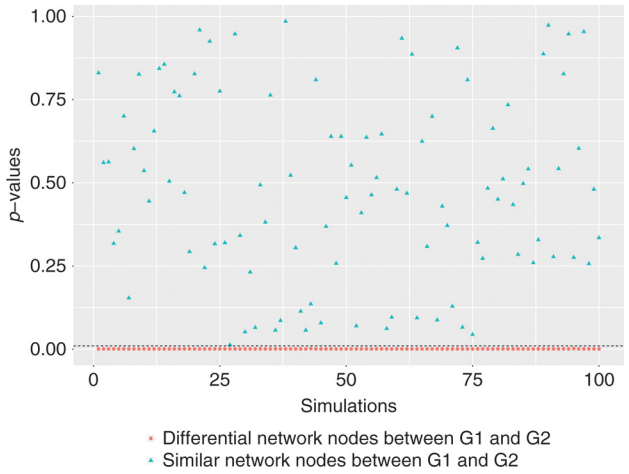


**Figure 1:** dNET *p*-values (x-axis) for 100 simulated datasets with (red dots) and without (blue triangles) differential effects. The black dotted line is the *p*-value cut off at 0.01.

### 3.1.3 Simulation study 2

This simulation study evaluated the effect of total sample size as well as the effect of the difference in group sample size on the sensitivity of dNET.

To analyse the effect of samples size, we simulated the datasets under scenario 2 for different sample sizes (ranging from 50 to 500). For each sample size, we generated 100 datasets with 77 genes, while keeping the group sizes same. The dNET results are given in Figure 2A where each boxplot displays the distribution of *p*-values (from 100 simulations) for each sample size. A large number of false negatives for datasets with sample size below 100 show that dNET require large sample sizes to detect moderate interaction effects with reasonable certainty.

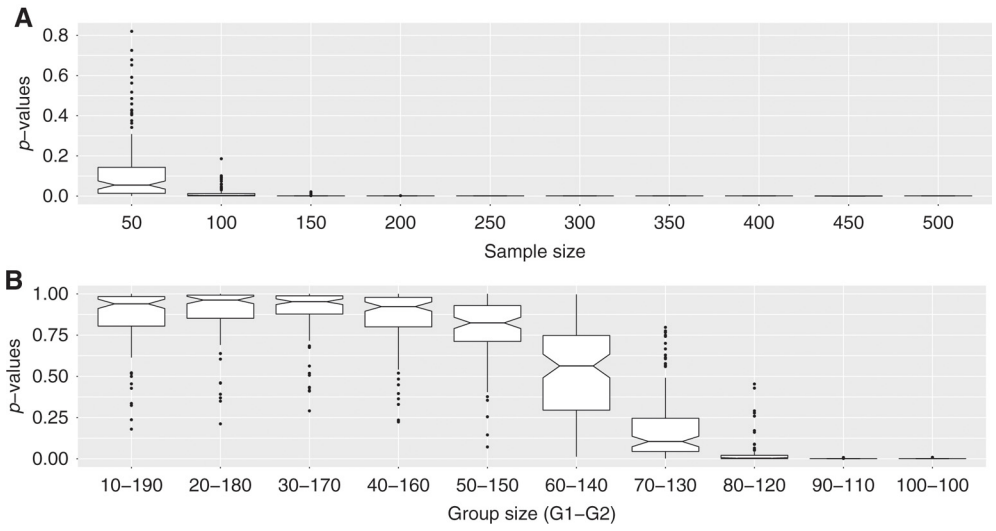


**Figure 2:** (A) dNET *p*-value distribution (y-axis) for different sample sizes (x-axis). (B) dNET *p*-value distribution (y-axis) for different group sizes (x-axis).

Similarly, for analysing the effect of the difference in group sample size we simulated datasets under scenario 2. In this case, we kept the total sample size same (200 samples) and varied the difference in group sample sizes from 180 ($n^{(a)}$ = 10 samples and $n^{(b)}$ = 190 samples) to 0 ($n^{(a)} = n^{(b)}$ = 100 samples). It is evident from the *p*-value boxplots in Figure 2B that a difference of 60 or more samples between two groups results in a large number of false negatives. Due to the small sample size and, probably, low power the interaction effects in the smaller group are harder to detect, therefore making it difficult for dNET to compare them across the two groups.

### 3.1.4 Simulation study 3

Our method, dNET, detects the differences in indirect trans effects between two groups of samples while integrating copy number and gene expression data. However, there can be instances when the associational differences between two groups of samples are at the level of direct cis-effects (gene-dosage led gene expression) instead of indirect trans effects. In this simulation study, we assess the performance of dNET by applying it to datasets where only the cis-effects differ. To do this, we simulated 100 datasets (77 genes and 200 samples) under case 1, case 2, and case 3, with no differences in the indirect trans effects between the two groups.

For comparing the results between different cases, we counted the number of dNET *p*-values larger than 0.05 (not significant) or smaller than 0.05 (significant) and plotted the total numbers under each case (Figure 3A, represented by the y-axis). Since the datasets do not exhibit any differential indirect trans effects and dNET is designed to detect only them, any significant *p*-value from this study was considered as false positive. It can be seen from the plot that under each case almost 95% of dNET *p*-values are not significant (true negatives) with only a few false positives. This shows that dNET is robust against differential cis-effects and targets only indirect trans effects.

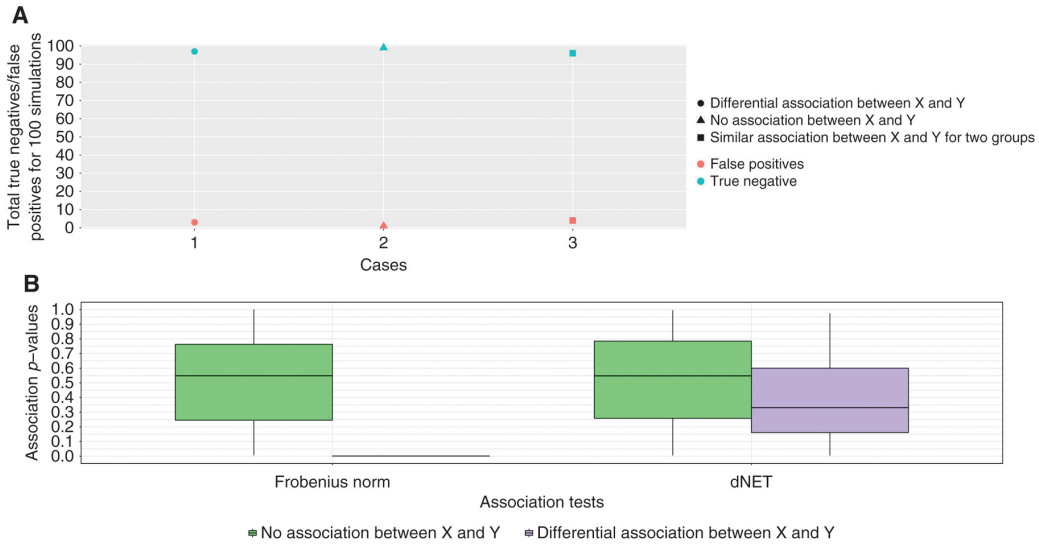


**Figure 3:** (A) Total number of false positives and true negatives (y-axis) for three different cases (x-axis). The number of false positives is shown by red colour and the number of true negatives is shown by blue colour. (B) *p*-Value boxplots from the naive approach based on Frobenius norm and dNET. The green boxplots represent the case with no cis-associations between X and Y. The purple boxplots represent the case with cis-associations between X and Y.

We also compared dNET with a naive approach which estimates the group-wise covariance matrices from the expression data only and then compare them using the Frobenius norm as test statistics, under a permutation scheme. For this, we simulated datasets under case 1 (no associations between **Y** and **X**) and case 2 (differential cis-effects between **Y** and **X**) and performed the analysis using the two approaches. The boxplots in Figure 3B show the analysis *p*-values from dNET and the naive approach based on Frobenius norm. As shown before, dNET does not detect any significant associations under case 1 and case 2. The Frobenius norm based approach performs well under case 1, with no significant associations detected. However, under case 2 it gives significant *p*-values. These results show that, unlike dNET, non-integrative approaches based on separate analysis of subgroups are not robust against differential cis-effects and hence are prone to false positives.

## 4 Application on TCGA cancer datasets

We apply dNET to bladder (n = 120, p = 16,612) and colon (n = 122, p = 12,423) TCGA cancer datasets. For each cancer we analysed 7 different signaling pathways, namely Apoptosis, ErbB, MAPK, Notch, PI3K-Akt, TGF-beta, Wnt, mTOR and p53 signaling pathways.

We consider the pathological stage of the samples as the grouping variable. Cancer staging describes the severity of an individual's cancer based on the magnitude of the original (primary) tumour as well as on the extent cancer has spread in the body. It has been shown to have strong associations with distinct gene expression patterns (Kheirelseid et al., 2013; Yao et al., 2015) and the prognosis of the disease. This makes cancer stage an interesting grouping variable when studying differences in signaling pathways in the TCGA cancer datasets.

We consider all patients with cancer stage lower than stage 3 as group *a* and all stage 3 or above patients as group *b*. The group size for each cancer is given in Table 1.

**Table 1:** Group size for TCGA cancer datasets.

| TCGA datasets | Cancer stage = stage 2 ($g_a$) | Cancer stage = stage 3 ($g_b$) |
|---|---|---|
| Bladder | 34 | 86 |
| Colon | 73 | 49 |

With dNET, we can test if there are differences in pathway topology between two groups of samples. As described in Section 2.4, we fit the model per gene for all samples together and then calculate the test statistic using the estimated parameter matrices. dNET produces a single *p*-value for every pathway giving an overview of the differences between the two groups of samples. Based on these *p*-values one can select the interesting pathway for gene level analysis for identifying individual pathway nodes. The fact that dNET considers sets of variables and produces a single *p*-value, reduces the need for multiple testing.

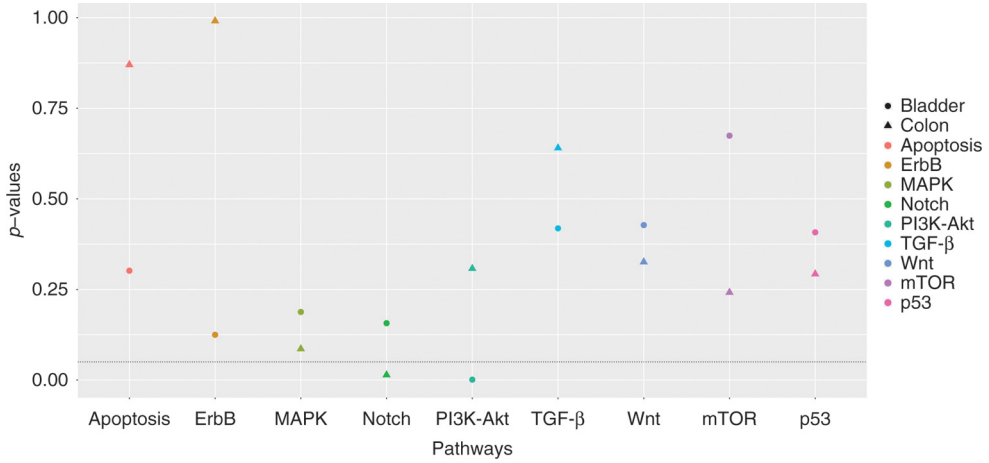


**Figure 4:** dNET *p*-values for different pathways in bladder and colon cancer. Each symbol represents a single *p*-value (y-axis) corresponding to the test for differential pathway topology (x-axis) between two groups of cancer (bladder or colon) samples. The horizontal dotted line denotes the *p*-value cut-off at 0.05.

Figure 4 displays the dNET *p*-values results for colon and bladder cancer for differences in pathway nodes between group *a* and group *b* samples. Despite the small sample sizes the method identifies pathways showing differential topology ($p < 0.05$) in both cancers. For bladder cancer, the dNET *p*-values for PI3K-Akt signaling pathway is significant ($p = 0.0009$) and for colon cancer, Notch signaling pathway has a significant dNET *p*-value ($p = 0.013$).

To understand the association patterns contributing towards these *p*-values, we estimated the $\boldsymbol{\Theta}$ coefficients by fitting the ridge model separately for the two groups of samples. By visualising the two coefficient matrices as heatmaps we can identify individual pathway nodes that are different between group *a* and group *b* samples. For example, Figure 5 and Figure 6 show these heatmaps for Notch signaling pathway in group *a* and group *b* colon cancer samples, respectively. From the figures, it can be seen that in colon cancer samples with stage 3 or above, expression of gene *HDAC*2 downregulates gene *DLL*1 with a negative $\theta$ value of $-0.40$. However, stage 2 or below colon cancer samples do not show this association and the $\theta$ value for copy number mediated regulation of gene *DLL*1 by *HDAC*2 is $-0.003$. This visualisation of the pathway topologies can be helpful for locating the individual effects contributing towards dNET test statistic and designing further analysis based on these observations. Similarly, blocks of genes, with small effect sizes, can also be spotted from the heatmaps showing differential topology between group *a* and group *b* samples. These differential effects might be too weak for individual testing but dNET test statistic benefits from these small, yet important, effects by taking them all together.

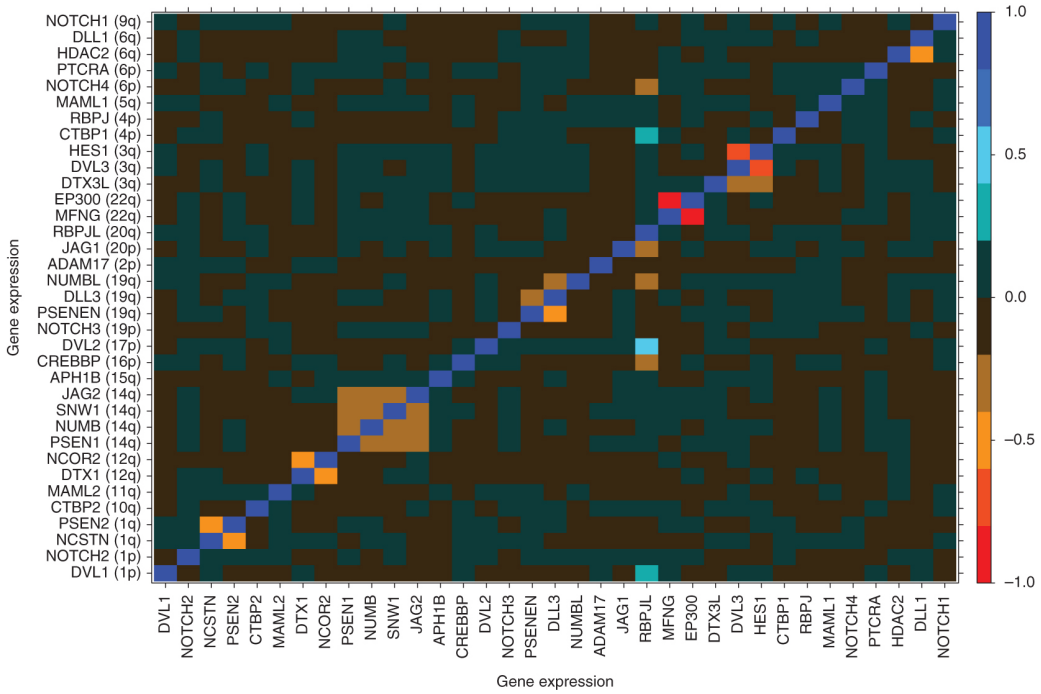


**Figure 5:** Heatmap showing the T values in Notch signaling pathway for all stage 3 or above colon cancer samples.

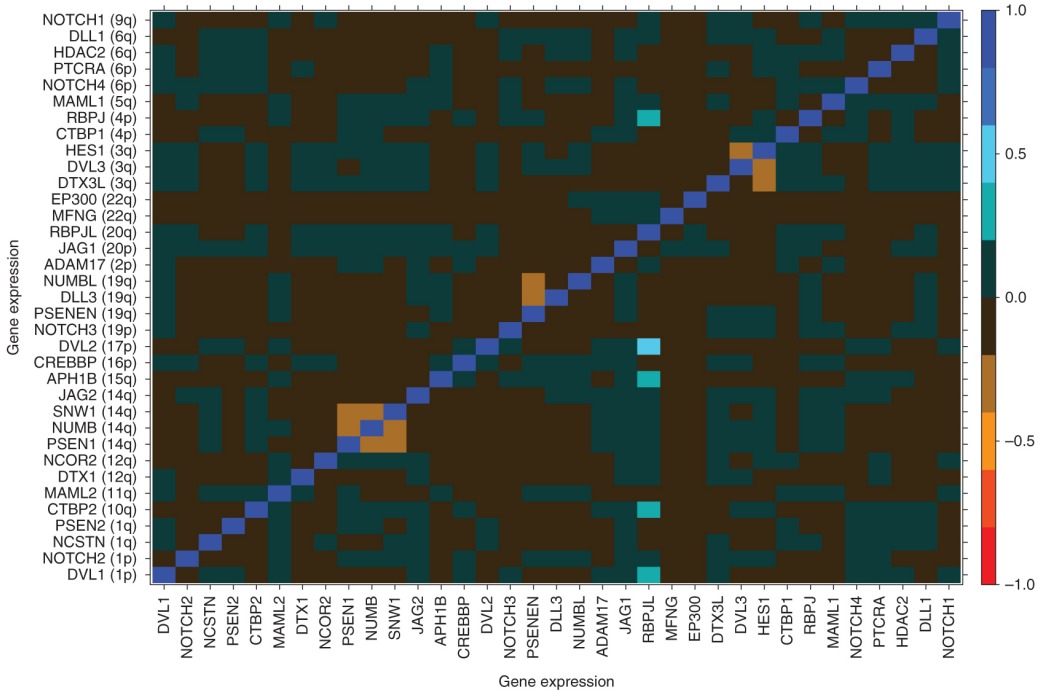


**Figure 6:** Heatmap showing the T values in Notch signaling pathway for samples with colon cancer stage lower than stage 3.

## 5 Discussion and conclusion

The work presented in this paper proposes a test for differential network analysis, named dNET, which integrates high dimensional genomic datasets from different molecular levels. It builds on recent research that uses rate equations for estimating indirect trans effects (van Wieringen & van de Wiel, 2014) by adding a multivariate test for differential group testing.

Instead of testing for groups separately, our method uses data from all samples at the same time in the model. This makes it less susceptible to problems arising due to small samples sizes, such as low power or artificial signals due to different group sample sizes. Moreover, dNET is designed to work with variable sets and test for

an overall difference between the network topologies for two groups of samples. Working with variable sets is advantageous over individual testing for several reasons. Firstly, it allows better detection of pathway nodes by using the biological information for all gene pairs together in a single model. Secondly, it allows detection of weak differential effects by combining information across related set of genes. Thirdly, by focusing on variable sets it reduces the number of tests and hence minimises the multiple testing burden. Finally, by making the analysis computationally less intensive and less time consuming it helps the data analyst to focus more quickly on the areas with interesting effects.

Although the test is designed for variable sets, the parameters for calculating the test statistic are estimated per gene. To take care of $p \gg n$ problem during estimation, dNET uses ridge regression model for obtaining these parameters. By using permutation testing, we ensure that the test is free of any parametric assumptions and gives an exact $p$-value.

Analysis of publicly available TCGA cancer datasets as well as simulated datasets shows that dNET can detect differential indirect trans effects with high accuracy and the presence or absence of differential cis-effects does not impact the $p$-value. However, like any other statistical method, dNET also relies on large sample sizes to detect these complex associations. From the application of dNET on publicly available TCGA cancer dataset, we not only detect pathways with differential topology between groups of samples but also reveal the possible contribution of individual pairs of genes, towards the overall dNET test statistic. These results show that separate group analysis, as the one suggested in (van Wieringen & van de Wiel, 2014), may complement dNET results by highlighting individual differential nodes in a pathway, although it can suffer from low power for detecting the pathway nodes within a group of samples, due to small sample size.

The implementation of the method and the analysis is done using R.

## Acknowledgement

This work has been supported and funded by Netherlands Bioinformatics Centre.

## References

Chaturvedi, N., J. J. Goeman, J. M. Boer, W. N. van Wieringen and R. X. d. Menezes (2014): "A test for comparing two groups of samples when analyzing multiple omics profiles," BMC Bioinformatics, 15, 1–14.

Chaturvedi, N., R. X. d. Menezes and J. J. Goeman (2017): "A global x global test for testing associations between two large sets of variables," Biometrical J., 59, 145–158.

Flutre, T., X. Wen, J. Pritchard and M. Stephens (2013): "A statistical framework for joint eqtl analysis in multiple tissues," PLoS Genet., 9, e1003486.

Goeman, J. J., S. van De Geer and H. van Houwelingen (2006): "Testing against a high dimensional alternative," J. R. Stat. Soc. Series B Stat. Methodol. 68, 477–493.

Hoerl, A. and R. Kennard (1970): "Ridge regression: biased estimation for nonorthogonal problems," Technometrics, 12, 55–67.

Horlings, H., C. Lai, D. Nuyten, H. Halfwerk, P. Kristel, E. van Beers, S. Joosse, C. Klijn, P. Nederlof, M. Reinders, L. Wessels and M. van de Vijver (2010): "Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients," Clin. Cancer Res., 16, 651–663.

Jornsten, R., T. Abenius, T. Kling, L. Schmidt, E. Johansson, T. E. M. Nordling, B. Nordlander, C. Sander, P. Gennemark, K. Funa, B. Nilsson, L. Lindahl and S. Nelander (2011): "Network modeling of the transcriptional effects of copy number aberrations in glioblastoma," Mol. Syst. Biol., 7, 486.

Kendziorski, C. M., M. Chen, M. Yuan, H. Lan and A. D. Attie (2006): "Statistical methods for expression quantitative trait loci (eqtl) mapping," Biometrics, 62, 19–27.

Kheirelseid, E., N. Miller, K. Chang, M. Nugent and M. J. Kerin (2013): "Clinical applications of gene expression in colorectal cancer," J. Gastrointest. Oncol., 4, 144–157.

Listgarten, J., C. Kadie and D. Heckerman (2010): "Correction for hidden confounders in the genetic analysis of gene expression," Proc. Natl. Acad. Sci. USA, 107, 16465–16470.

Menezes, R. d., M. Boetzer, M. Sieswerda, G. van Ommen and J. M. Boer (2009): "Integrated analysis of dna copy number and gene expression microarray data using gene sets," BMC Bioinformatics, 10, 203.

Phipson, B. and G. K. Smyth (2010): "Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn," Stat. Appl. Genet. Mol. Biol., 9, 1544–6115.

van Iterson, M., S. Bervoets, E. de Meijer, H. Buermans, P. 't Hoen, R. d. Menezes and J. Boer (2013): "Integrated analysis of microrna and mrna expression: adding biological significance to microrna target predictions," Nucleic Acids Res., 41, 1–10.

van Wieringen, W., J. Berkhof and M. van de Wiel (2010): "A random coefficients model for regional co-expression associated with DNA copy number," Stat. Appl. Genet. Mol. Biol., 9, 1–28.

van Wieringen, W. N. and M. A. van de Wiel (2014): "Penalized differential pathway analysis of integrative oncogenomics studies," Stat. Appl. Genet. Mol. Biol., 13, 141–158.

Yao, F., C. Zhang, W. Du, C. Liu and Y. Xu (2015): "Identification of gene-expression signatures and protein markers for breast cancer grading and staging," PLoS One, 10, 1–17.