MetaDCN: meta-analysis framework for differential co-expression network detection with an application in breast cancer

Li Zhu1,†, Ying Ding2,†, Cho-Yi Chen1,3,†, Lin Wang1, Zhiguang Huo1, SungHwan Kim1, Christos Sotiriou4, Steffi Oesterreich5 and George C. Tseng1,2,*

1Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA, 2Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA, 3Genome and Systems Biology Degree Program, National Taiwan University, Taipei 10617, Taiwan, 4Breast Cancer Translational Research Laboratory, J. C. Heuson, Institut Jules Bordet, University Libre de Bruxelles, Brussels 1000, Belgium and 5Magee-Women’s Research Institute, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors
Associate Editor: John Hancock

Received on September 14, 2016; revised on November 11, 2016; editorial decision on December 2, 2016; accepted on December 7, 2016

Abstract

Motivation: Gene co-expression network analysis from transcriptomic studies can elucidate gene–gene interactions and regulatory mechanisms. Differential co-expression analysis helps further detect alterations of regulatory activities in case/control comparison. Co-expression networks estimated from single transcriptomic study is often unstable and not generalizable due to cohort bias and limited sample size. With the rapid accumulation of publicly available transcriptomic studies, co-expression analysis combining multiple transcriptomic studies can provide more accurate and robust results.

Results: In this paper, we propose a meta-analytic framework for detecting differentially co-expressed networks (MetaDCN). Differentially co-expressed seed modules are first detected by optimizing an energy function via simulated annealing. Basic modules sharing common pathways are merged into pathway-centric supermodules and a Cytoscape plug-in (MetaDCNExplorer) is developed to visualize and explore the findings. We applied MetaDCN to two breast cancer applications: ER+/ER- comparison using five training and three testing studies, and ILC/IDC comparisons with two training and two testing studies. We identified 20 and 4 supermodules for ER+ and ILC/IDC comparisons, respectively. Ranking atop are ‘immune response pathway’ and ‘complement cascades pathway’ for ER comparison, and ‘extracellular matrix pathway’ for ILC/IDC comparison. Without the need for prior information, the results from MetaDCN confirm existing as well as discover novel disease mechanisms in a systems manner.


Contact: ctseng@pitt.edu

Supplementary information: Supplementary data are available at Bioinformatics online.
1 Introduction

Differential co-expression (DC) refers to the change in gene–gene correlations between two conditions (e.g. cases and controls). Changes in gene–gene correlation may occur in the absence of differential expression, meaning that a gene may undergo radical changes in regulatory pattern that would be undetected by traditional differential expression (DE) analyses (see Fig. 1A). A specific phenotype could be contributed by differential co-expression without altering the expression levels of genes. This phenomenon has been found in aging (Southworth et al., 2009) as well as other biological conditions (Gaiteri et al., 2014). Disease-associated alterations in the regulatory systems that create co-expression changes may be revealed through comparing gene–gene correlations that are computed separately from control and disease populations. Therefore, DC analysis can provide complementary information to standard differential expression (DE) analyses. Differential co-expression in two conditions could shed light on novel biological mechanism. For example, a group of genes may be regulated by a common transcription factor or epigenetic modification, which is active in one condition but disrupted in the other.

In the literature, Lai et al. (2004) has proposed an expected conditional F-statistics to identify differential co-expressed gene pairs, while Amar et al. (2013) and Bhattacharyya and Bandopadhyay (2013) developed methods for direct identification of DC gene modules. Choi and Kendziorski (2009) detected differential co-expression using predefined gene sets such as Gene Ontology (GO) categories. Although this approach incorporates prior biological information, it lacks the ability to detect novel DC modules. Another class of methods detected differential modules with genes highly co-expressed in one reference condition but with little or no correlation in the other condition. These types of methods rely on applying clustering methods to one reference condition, causing case-control asymmetry in the analysis (Imhels et al., 2005; Watson, 2006). To circumvent this problem, Zhang and Horvath (2005) identified co-expressed modules in the entire (cases and controls combined) cohort through clustering and then evaluate their differential co-expression across conditions. Similarly, Tesson et al. (2010) extended this framework to detect differential co-expression modules by introducing the correlation changes between conditions into dissimilarity matrix for clustering (DiffCoEx).

All methods described above for DC network detection focused on single transcriptomic study analysis. The differential correlation relationship could arise from meaningful biological sources as well as uncorrected technical biases (see Figure 1 in Gaiteri et al., 2014).

Any mechanism that synchronously regulates transcription of multiple genes, unwanted batch effect, or mixture of tissues could potentially contribute to co-expression relationships. Therefore, instead of looking for DC networks between two conditions in a single study, differential co-expression may be confirmed across multiple datasets via meta-analyses to increase the detection power and stability. DC networks that are significant in one dataset may become more convincing if the DC patterns are preserved across multiple datasets. DC between conditions can be assessed by different choices of measures; for example, differential modules with a predominant measure such as density (Li et al., 2011) or other sophisticated network measures (Kugler et al., 2013; Langfelder et al., 2011).

So far, few studies attempted to detect DC networks across multiple studies. Mehan et al. (2009) proposed a simulated-annealing-based method to detect DC modules of which the network density changes were associated with phenotype. However, their method embedded pathway enrichment in the optimization of objective function; that is, the optimization phase heavily depended on the prior knowledge and also the output module sizes from the method were generally small. In this paper, we have developed a new meta-analytic framework, namely MetaDCN, to search for initial DC modules without prior information. Our method included additional network properties in the energy function to detect biologically meaningful ‘basic DC modules’ and false discovery rate (FDR) was controlled by permutation analysis. We then further combined basic DC modules that share common pathway annotation into more interpretable DC supermodules. We evaluated the method on simulated data and breast cancer studies to search for DCN between ER + versus ER- and invasive lobular carcinoma (ILC) versus invasive ductal carcinoma (IDC). The identified DCN were further validated in independent breast cancer studies. The result identified pathways such as ER-mediated immune functions and extracellular matrix heterogeneity between ILC and IDC that help elucidate the underlying disease mechanisms.

2 Methods

MetaDCN combines multiple case-control transcriptomic studies to detect disease-associated modules such that genes in the modules are highly correlated in control samples but the correlations are...
disrupted in cases or vice versa. An energy function is introduced to
detect modules of DC networks across studies. Since direct optim-
ization for large modules is computationally challenging and un-
stable, we will first aim for detecting a sequence of small ‘basic DC
modules’ of sizes between 3 and 30. Basic DC modules are then
combined into DC supermodules via a module assembly algorithm
based on pathway enrichment information (see Fig. 1B). Such
pathway-centric assembly improves functional annotation of de-
tected supermodules that can advance disease understanding and
guide further hypothesis generation.

2.1 Basic DC module detection
The algorithm to detect basic DC modules is outlined below. Details
of energy function, optimization procedure and false discovery rate
control are described.

2.1.1 Energy function
Consider N transcriptomic studies, each containing case and control
samples. Gene co-expression networks are first constructed among
cases and among controls for each of the N studies, thus generating
2N co-expression networks. In this paper, we demonstrate our
method based on unweighted networks but the method can be ex-
tended to weighted networks. To build unweighted networks and
normalize them across different studies, we first calculated pair-wise
gene–gene Spearman’s correlations for robust comparisons. In con-
trast to Pearson correlation, Spearman’s correlation can capture
both linear and non-linear association. Considering the large num-
ber of possible edges and computation complexity, we then select
the correlation cut-off for edge connections so that only the top
0.4% of possible connections in each network were kept (Lee et al.,
2003). This procedure provides robustness because different studies
usually have different sample sizes and are conducted using different
experimental platforms, which could result in distinct correlation
distributions. Our proposed algorithm is developed for the more
popular unweighted network but it can be modified for weighted
network if desired.

We propose to minimize the following energy function (target
function) for detection of gene modules with differential co-
expression:

\[ E_{\text{tot}} = w_1 E_{\text{diff-mean}} + w_2 E_{\text{size}} + w_3 E_{\text{diff-var}} \]

The proposed target function comprises the following three compo-
nents: I) \( E_{\text{diff-mean}} \) for mean network density difference between two
phenotypes across N studies, II) \( E_{\text{size}} \) for size of module and III)
\( E_{\text{diff-var}} \) for the consistency of the density difference between the
two phenotypes across N studies. Gene modules minimizing \( E_{\text{tot}} \) have
consistently large correlation difference between cases and controls
across multiple studies, as well as reasonable large size. The search
direction is bidirectional, meaning that we will identify modules
with significantly higher connections in case networks than in con-
rol networks and then repeat reversely.

Each component in the target function is described by an expo-
ential decay function. The first component is defined as

\[ E_{\text{diff-mean}} = \exp \left\{ -z_1 \left( \frac{N}{X} \sum_{i=0} \left( \delta_{i} \cdot \delta_{i+1} \right) \right) \right\} \]

where \( \delta_{i} \) and \( \delta_{i+1} \) are the densities of case network and control network
respectively in study \( i \). The exponential decay function favors
larger mean density differences between cases and controls and is
the major target of our algorithm. The second component
\( E_{\text{size}} = \exp \left\{ -z_2 \left( \frac{\mid x \mid}{\gamma} \right) \right\} \) (where \( x \) denotes the genes in the module,
\( \mid x \mid \) is the module size) is related to the size of the modules which
favors larger modules and penalizes smaller modules. We restricted
the module size no larger than 30 due to large searching space. We
set \( \gamma = 30 \) to re-scale the ratio ranging from 0 to 1 to make the
three decay parameters (\( z_1 \), \( z_2 \) and \( z_3 \)) comparable in later
parameter selection. Without \( E_{\text{size}} \), dimers or triplets with density 1
or 0 could easily dominate the output by random chance and in-
crease false positives. The third component \( E_{\text{diff-var}} = \exp \left\{ -z_3 \left( 1 - V \right) \right\} \]
quantifies the variance of the paired difference of network densities between cases and controls
across studies to favor consistent differential co-expression among
studies. In all three components, the parameters \( z_1 \), \( z_2 \) and \( z_3 \) con-
trol the decay rate in the exponential function. In our implementa-
tion, we set \( z_2 = 10 \) for \( E_{\text{size}} \) and \( z_1 \) and \( z_3 \) = 5 for \( E_{\text{diff-mean}} \)
and \( E_{\text{diff-var}} \). The higher \( z_2 \) was used for \( E_{\text{size}} \) to avoid extremely small
modules as previously mentioned.

To tune the parameters \( w_1 \), \( w_2 \) and \( w_3 \) in the target function, we
first constrain the sum of the three parameters to be 1000, i.e.
\( w_1 + w_2 + w_3 = 1000 \). We assigned equal importance to \( E_{\text{diff-mean}} \)
and \( E_{\text{diff-var}} \) by setting \( w_1 = w_2 \) and searched for optimal \( w_3 \) from
100 to 700 with 100 increments that could output the largest
number of basic DC modules under FDR 0.3 (see below for detection
of basic DC modules and FDR control).

2.1.2 Optimization by simulated annealing
Due to the non-convex nature of \( E_{\text{tot}} \), we applied simulated anneal-
ing, a stochastic algorithm for non-convex optimization
(Kirkpatrick et al., 1983). In each Monte Carlo (MC) step with
simulated annealing, a new state is proposed and denoted as \( \text{X}_{\text{new}} \),
which is either adding a node (gene) from trial set to selected set or
removing a node (gene) from selected set to trial set. At the begin-
ing, the trial set is determined as the set of genes that have at least
one edge connected to the seed module genes (initial selected set) in
any of the N case co-expression networks. If the resulting energy is
smaller, the state is accepted. If not, the state is accepted with an
acceptance probability as \( P_{\text{acc}} = \min \left\{ 1, \frac{\pi(x_{\text{new}}}{{\pi(x_{\text{old}}} \right) \right\} \right\}$
where \( P_{\text{acc}} \) is the acceptance probability and \( \pi(x_{\text{new}}} \) is the transition prob-
ability from old state to the new state. If a genes is added from trial
set to selected set, \( P(x_{\text{old}}-\text{new}} \) for \( x_{\text{old}} \) to \( x_{\text{new}} \) follows a Bernoulli dis-
tribution, where \( |x| \) denotes the size of set \( x \). \( \pi(x_{\text{new}}} \) is the Boltzmann
distribution of the energy function to be minimized: \( \pi(x_{\text{new}}} \) = \exp \left\{ -E_{\text{tot}} \left( x_{\text{new}} \right) / T \right\}$
where \( T \) is a temperature parameter. When temperature is high, new trial
moves will be accepted easily and thus more freely jump out of the
local minimum. When temperature gets lower, it tends to converge
to a local minimum. We apply the temperature schedule \( T(k+1) = 0.95 \cdot T_k \)
and stop the annealing run if the acceptance ratio is smaller than 2%,
where the acceptance ratio is calculated as the ratio of steps accepted in every 400 MC steps. Due to large search-
ing space, we bounded the module size between 3 and 30. If current
module size is 3, only addition of new node is allowed for a new
state while if module size is 30, only node removal is allowed.

Although simulated annealing helps improve local minimum
trapping, a good starting point, which is called seed module here, is
critical for optimization in high dimensional space. Instead of ran-
domly selecting a subset of genes from the genome to be the seed
modules, an edge-study matrix of Spearman correlations was con-
structed where rows represent all possible edges and columns repre-
sent all studies in two conditions of size 2N (Walley et al., 2012).
For each edge on the rows, a simple paired t-test is applied to the
Spearman correlations to assess candidate differential co-expression edges (require paired $t$-test $P$-value < 0.1 and absolute mean difference of Spearman correlation > 0.1). Based on these candidate differential co-expression edges, an initial network is constructed and multiple (denoted as $K$) connected graphs in the network are identified. If the size of a connected graph is larger than 30, we randomly sample 10 genes from it as the initial seed module for optimization starting points; if the size is smaller than 3, we discard it. Otherwise, the optimization starts from the connected graphs as the seed module directly. In our evaluation for simulations and application, such an algorithm to generate seed modules has performed well. But it is possible to apply other community detection algorithms for this purpose (Fortunato, 2010).

Although simulated annealing helps improve local optimum problem, optimization instability still exists. We will repeat the optimization by starting from $K$ initial seed modules and repeat $K$ times of simulated annealing repeats. For two repeats with Jaccard index greater than 0.8, we will select the one with smaller $E_{\text{tot}}$. This will generate $\sum_{k=1}^{K} R_k$ basic differential co-expression (DC) modules for supermodule assembly, where $R_k$ is the number of basic modules from the $k$th seed module with pairwise Jaccard index smaller than 0.8.

### 2.1.3 Control of false discovery rate
To avoid detection of spurious modules by chance, false discovery rate is assessed for detected $\sum_{k=1}^{K} R_k$ basic DC modules as described below. Denote by $E_{b_k}$ the optimized energy value for detected basic DC module $u_{b_k}$ from the $k$-th seed module and $j$-th simulated annealing repeat, where $1 \leq k \leq K$ and $1 \leq j \leq R_k$. We first permute the case-control class labels for samples in each study and then reconstruct the case and control co-expression network as described previously. Simulated annealing optimization is similarly applied to detect $\sum_{k=1}^{K} R_k$ 'null' basic DC modules, where $R_k$ is the number of basic modules detected from permuted network with pairwise Jaccard index smaller than 0.8. Suppose the permutation is repeated for $B$ times and the resulting energy values are denoted as $E_{b_k}^B$, where $1 \leq b \leq B, 1 \leq k \leq K^B, 1 \leq j \leq R_k^B$. Under null hypothesis, the resulting case and control co-expression networks from permutation have no difference and $E_{b_k}^B$ will form a null distribution to assess $P$-values of $E_{b_k}$. The $P$-values of basic DC modules $u_{b_k}$ are estimated as

$$p(u_{b_k}) = \frac{\# \{ s_{ij} \mid \sum_{j=1}^{B} \sum_{k=1}^{K} \sum_{j=1}^{R_k} I(E_{b_k}^s \leq E_{b_k}^B) \leq 1 \} + 1}{\sum_{j=1}^{B} \sum_{k=1}^{K} \sum_{j=1}^{R_k} 1 \} + 1},$$

where $\# \{ s_{ij} \mid \sum_{j=1}^{B} \sum_{k=1}^{K} \sum_{j=1}^{R_k} I(E_{b_k}^s \leq E_{b_k}^B) \leq 1 \}$ is the number of basic modules detected from permuted network with pairwise Jaccard index smaller than 0.8.

### 2.2 Supermodule assembly, summarization and visualization

#### 2.2.1 DC supermodule assembly
Since the current approach limits the size of the basic DC modules between 3 and 30, small modules often do not yield significant pathway enrichment annotation to inspire further hypothesis generation. Therefore, in order to obtain larger DC modules, we proposed to use statistical significance of pathway enrichment to guide module assembly. Firstly, we applied pathway enrichment analysis using Fisher’s exact test on detected basic DC modules (here we choose FDR $\leq 0.3$) against 2,379 pathways downloaded from MSigDB (http://www.broadinstitute.org/gsea/msigdb/), which contained Biocarta, KEGG, Reactome and Gene Ontology databases (excluding large pathways with more than 250 genes). For each given pathway, we applied Fisher’s meta-analysis method to combine $P$-values across basic DC modules and selected the top 150 pathways with the most significant meta-analyzed $P$-values. The restriction is not necessary, but will reduce the computation cost, without changing the results much. For each of the 150 candidate pathways, we searched among combinations of up to three basic DC modules (including both over-connected and under-connected DC modules of case-control comparison) and identify the assembled supermodule such that its pathway enrichment $P$-value is minimized. Take the immune response pathway in Figure 2C as an example, the pathway enrichment $P$-values for modules H9, L1 and L2 (H stands for modules with higher density in ER+ patients; L stands for modules with lower density in ER+ patients) are 0.018, $7 \times 10^{-4}$ and 0.02 with module sizes 10, 10 and 11, respectively. The supermodule combining these three basic DC modules contains 28 genes with Fisher’s exact test $P$-value = $1.33 \times 10^{-6}$. Assembly of multiple basic DC modules can yield larger supermodules with more genes involved in a specific pathway, which provides better biological interpretation and hypothesis generation. Additionally, if the assembled supermodule contains both over-connected and under-connected basic DC modules (see red and green edges in Fig. 2C), it may suggest an interesting alternative activation mechanism in the pathway related to disease development.

#### 2.2.2 Summarization and visualization of DC supermodules
Visualization of basic DC modules across $N$ studies can be easily done by displaying the $2N$ co-expression networks as shown in Figure 2(A, B). For DC supermodules, however, smarter design of visualization is needed. Figure 2(C, D) shows our proposed visualization display for DC supermodules. On the left plot, three basic DC modules (gene nodes displayed by red, blue and yellow) are combined to form the DC supermodule. The edge widths between any pair of gene nodes $i$ and $j$ are controlled proportionally by a standardization score $Z_{ij}$ to represent the degree of differential co-expression. Denote by $u_{ij}^{(s)}$ and $v_{ij}^{(s)}$ the Spearman correlation between gene $i$ and $j$ in study $s$ in case and control samples, respectively. Let $d_{ij}^{s} = u_{ij}^{(s)} - v_{ij}^{(s)}$ and let $d_{ij}$ be the mean of paired correlation differences of all studies, $\sigma_d$ be the standard deviation of paired correlation differences and $\sigma_d$ be the fudge parameter estimated by the median of all standard deviations $\sigma_d$’s. The edge widths are proportionally to the standardized score $Z_{ij} = \frac{d_{ij}}{\sigma_d}$. The fudge parameter $\sigma_d$ is to avoid accidentally large $Z_{ij}$ due to small $\sigma_d$ (Tusher et al., 2001). As a result, the DC supermodule can be represented as a weighted undirectional network. $P$-values of the $Z$ scores were calculated by permuting case and control subjects in each study and randomly subsample the same number of genes to calculate the null permuted $Z$ scores and comparing with them. Only edges with significant $P$-values passing certain $P$-value threshold are displayed in the network plot.

We further developed a Cytoscape plug-in application, called ‘MetaDCExplorer’, which utilizes the power of Cytoscape Java API in visualizing complex networks and integrating topology with attributes. The interface allows users to load the input supermodule attributes and generate interactive network visualization with additional context annotations. First, the user selects a DC supermodule of interest to visualize from the list of biological pathways ranked by the significance of enrichment. The attributes of that supermodule will be loaded. The absolute values of the standardized $Z$ scores for each gene pairs will be interpreted as edge widths, and the initial network view will be generated using edge-weighted force directed layout algorithm provided from prefuse toolkit (see supplement
In this paper, we applied MetaDCN to two breast cancer applications. The first application detects DC supermodules for biological hypotheses on potential disease mechanisms. In the second application, DC supermodules are detected for sub-pathways that allow biologists to further explore and generate new hypotheses.

### 2.3 Datasets

In this paper, we applied MetaDCN to two breast cancer applications. In the first application, DC supermodules are detected for ER+ versus ER- comparison in five training studies and validated in two testing studies, where only ILC subjects are available. Details of data description and data preprocessing are available in Supplement Material.

### 3 Results

#### 3.1 Simulation

We first applied MetaDCN to a simulated dataset including 5 studies. Each study contained case and control groups, with number of subjects in each group drawn from Poisson(50). We generated 1000 artificial genes named 1 to 1000 and a subset of them belonged to 5 gene modules (non-overlapping), each of which containing the top 100 most differentially expressed genes. We set $c = 2$ for cases, $s = 1, 2, \ldots, 5$, and $m = 1, 2, \ldots, 5$. We set different $(\rho_{m1}, \rho_{m2})$ pairs for five modules to include both strong and weak signals. They were set to be (0.3, 0.1), (0.1, 0.3), (0.5, 0.1), (0.1, 0.5) and (0.7, 0.1) for 5 modules respectively. Therefore, the first and second modules have smaller signals, while the fifth module has the strongest signal. For genes outside the module, the expressions were i.i.d. drawn from $N(0, 1)$.

---

**Fig. 2.** (A) Example module (L1) more densely connected in ER+ group with red nodes indicate genes belonging to the immune response pathway. (B) Example module (H7) more densely connected in ER- group with red nodes indicate genes belonging to the complement cascade pathway. Nodes represent genes and links between them represent co-expression relationship. Each column corresponds to one independent study. (C) Visualization of immune response pathway supermodules. The edge color represents the direction of differential gene co-expression, in which positive values (red color) represent ER-positive-favored co-expression and negative values (green color) show ER-negative-favored co-expression. Node color represents the origin of sub-modules, and the genes annotated under the immune response pathway are highlighted with dark circles. Edge width represents edge weight (Z score of differential co-expression) (Color version of this figure is available at Bioinformatics online.)

---

In the network view, the edge width represents the edge weight. Nodes with their neighbors connected by high-weight edges will be automatically clustered together, so that the modular organization will be revealed. The edge color represents the direction of differential co-expression interpreted from edge Z scores, in which positive values (red color) indicate over-connected edges and negative values represent under-connected edges in case-control comparison. Node color represents the original basic DC modules where the gene belongs, and the genes annotated under the selected biological pathway are highlighted with outer black circles. To account for the fact that different diseases might have different range of differential co-expression signals, we thus introduced additional factors that control the repelling and attracting force between and within the modules. These factors, together with the edge $P$-value cut-off threshold, are made adjustable in a control panel so that users can update the network view in real time. In summary, this application is designed to reveal the modular organization of DC supermodules and to suggest alternatively activated sub-pathways that allow biologists to further explore and generate biological hypotheses on potential disease mechanisms.
With this simulated dataset, we constructed the edge-study matrix based on Spearman correlation. The module search was performed using simulated annealing algorithm with maximum iterations as 500. \( R = 3 \) trials with different initial seed modules were repeated, and \( P \)-value was calculated using \( B = 10 \) permutations. In the end, the best module among 3 repeats was selected based on optimal \( P \)-value and energy. For simplicity, here we only evaluate performance of basic DC modules without module assembly. If the Jaccard index (ratio of the intersection set divided by the union set) of the identified basic DC module to the underlying truth is greater than 0.5, we denote this searching as a successful hit.

We generated 50 datasets and compared the performance of MetaDCN with an existing method DiffCoEx (Tesson et al., 2010). The implementation of DiffCoEx used the R code directly from the original paper with the default setting. The soft threshold, as the most sensitive tuning parameter, was chosen based on scale free topology fit (Zhang and Horvath, 2005). The hierarchical tree was cut using dynamicTreeCut R package (Langfelder et al., 2008).

Table 1 shows the lower and upper quartile of number of detected modules and the percentage of successful hit for each of the five modules under different FDR cut-offs for permutation test in the 50 repeated simulations. The result shows that DiffCoEx tends to detect many false positives while still miss the underlying true DC netowrks.

### Table 1. Percentage of successful hit (Jaccard index >0.5) in simulation study (50 repeats)

<table>
<thead>
<tr>
<th>Method</th>
<th>FDR</th>
<th>Upper and lower quartile</th>
<th>M1 (%)</th>
<th>M2 (%)</th>
<th>M3 (%)</th>
<th>M4 (%)</th>
<th>M5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetaDCN</td>
<td>0.1</td>
<td>(3, 5)</td>
<td>56</td>
<td>58</td>
<td>96</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>(4, 5)</td>
<td>72</td>
<td>74</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>(5, 5)</td>
<td>78</td>
<td>82</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>DiffCoEx</td>
<td>–</td>
<td>(3, 39)</td>
<td>8</td>
<td>8</td>
<td>30</td>
<td>26</td>
<td>37</td>
</tr>
</tbody>
</table>

### Table 2. Top pathway-centric supermodules with at least 3 pathway overlapping genes (with 10 repeats with different initial modules)

#### (A) Pathway name (ER+ versus ER-)

<table>
<thead>
<tr>
<th>Pathway name (ER+ versus ER-)</th>
<th>Pathway size</th>
<th>Module size</th>
<th># pathway genes</th>
<th>q-value</th>
<th>P-value</th>
<th>Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>REACTOME_COMPLEMENT_CASCADE</td>
<td>32</td>
<td>25</td>
<td>4</td>
<td>2.14E-05</td>
<td>1.93E-07</td>
<td>H7,H8</td>
</tr>
<tr>
<td>GO_IMMUNE_RESPONSE</td>
<td>235</td>
<td>28</td>
<td>7</td>
<td>5.63E-05</td>
<td>1.33E-06</td>
<td>H9,L1,L2</td>
</tr>
<tr>
<td>REACTOME_REGULATION_OF_COMPLEMENT_CASCADE</td>
<td>14</td>
<td>25</td>
<td>3</td>
<td>5.63E-05</td>
<td>2.47E-06</td>
<td>H7,H8</td>
</tr>
<tr>
<td>GO_ORGAN_MORPHOGENESIS</td>
<td>144</td>
<td>35</td>
<td>6</td>
<td>5.63E-05</td>
<td>2.80E-06</td>
<td>H3,H5,L9</td>
</tr>
<tr>
<td>BIOCARTA_TCYTOTOXIC_PATHWAY</td>
<td>14</td>
<td>23</td>
<td>3</td>
<td>5.63E-05</td>
<td>2.85E-06</td>
<td>H3,L5</td>
</tr>
</tbody>
</table>

#### (B) Pathway name (ILC versus IDC)

<table>
<thead>
<tr>
<th>Pathway name (ILC versus IDC)</th>
<th>Pathway size</th>
<th>Module size</th>
<th># pathway genes</th>
<th>q-value</th>
<th>P-value</th>
<th>Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO_PROTEASE_INHIBITOR_ACTIVITY</td>
<td>41</td>
<td>27</td>
<td>3</td>
<td>0.003</td>
<td>6.13E-05</td>
<td>L2,L4,L8</td>
</tr>
<tr>
<td>GO_PROTEINACEOUS_EXTRACELLULAR_MATRIX</td>
<td>98</td>
<td>15</td>
<td>3</td>
<td>0.003</td>
<td>0.00085</td>
<td>L5,L7</td>
</tr>
<tr>
<td>GO_EXTRACELLULAR_MATRIX</td>
<td>100</td>
<td>15</td>
<td>3</td>
<td>0.003</td>
<td>0.00085</td>
<td>L5,L7</td>
</tr>
</tbody>
</table>

Module starts with \( H \) indicates it is more densely connected in ER+ or ILC network, while module starts with \( L \) indicates it is more densely connected in ER- or IDC network.

3.2 Breast cancer studies (ER+ versus ER-)

We next applied our method to identify differentially co-expressed modules between networks from ER+ patients and networks from ER- patients. Estrogen receptor, indicating the cancer cell response to hormone estrogen, is an important marker in breast cancer cases for treatment selection. Detecting differential co-expression network between ER+ and ER- patients can help us better understand the difference of disease mechanism, thus designing specific therapies for ER+/ER- patients. In the analysis of training data, five pairs of gene co-expression networks were constructed for ER+ patients and ER- patients across the five studies. Edge-study matrices were calculated and connected components were obtained as initial seed modules for simulated annealing algorithm. FDR was calculated for each of the modules with \( B = 10 \) permutations. The best weights were selected based on the results from first 3 repeats with different initial modules. With the optimal weights and \( R = 10 \) repeats, at FDR \( \leq 0.3 \), 12 basic DC modules were detected as over-connected in ER+ networks while another 12 basic DC modules were detected as over-connected in ER- networks. Two example modules, one densely connected in ER+ networks and one densely connected in ER- networks, are illustrated in Figure 2(A, B). Both modules achieved FDR 0.02.

We tested varying number of repeats \( (R) \) in each seed module and the results of pathway-centric assembly are quite consistent. We identified 20 supermodules engaged in 40 pathways, sharing at least 3 overlapping genes with the enriched pathway. The top pathways associated with the assembled modules were listed in Table 2(A) (see full list in Supplementary Table S4). Among the list of summarized DC supermodules, ‘complement cascade pathway’ was with highest significance followed by ‘immune response pathway’. Figure 2(C, D) showed the network view for these two DC supermodules.

In the literature, studies have shown that estrogen receptors can regulate innate immune cells (Kovats, 2015). Cunningham and Gilkeson (2011) found ERs have prominent effects on immune function in both the innate and adaptive immune responses. ER\(_{x}\) expression is associated with outcome in patients with autoimmune diseases such as lupus. Possible alternative activations of immune and complement pathway between ER+ and ER- breast cancer patients have also been revealed in several research studies. Teschendorff et al. (2007) found that the heterogeneity in clinical outcome of ER- breast cancer patients is related with complement and immune pathway, while this association is not observed in ER+ patients.

We next validated those two supermodules in leave-one-out cross-validation (LOOCV). Each time, we left one study out as
testing set and used the remaining four studies as training set to perform module searching and module ensemble. In each LOOCV, 2 or 3 basic DC modules were merged into a DC supermodule in each pathway. We calculated network averaged densities in each basic DC module in the four training studies (on the left of dashed line) as well as the testing study (on the right of dashed line) in Figure 3(A) and Supplementary Figure S1(A). Similarly, box-plots of Spearman correlation distributions are plotted in Figure 3(B) and Supplementary Figure S1(B). The result consistently shows good validation of the finding.

Finally, we used the top two pathways and the DC supermodules obtained from five training studies and tested in the three independent validation studies. Same set of genes was used for constructing co-expression network. If genes were not available in a study with different platform, the overlapped gene set was used. Following Figure 3 (A, B) and Supplementary Figure S1(A, B) for LOOCV, we plotted the average network densities and box-plots of Spearman correlation distribution in Figure 3(C, D) and Supplementary Figure S1(C, D) for the basic DC modules of the supermodules enriched in those two pathways. The result provides consistent validation of the differential co-expression pattern of gene modules enriched in these pathways.

As a comparison, we also applied DiffCoEx (Tesson et al., 2010) to our datasets. Since DiffCoEx is only applicable to a single study, we applied it to the largest study METABRIC using the same procedure as described in the simulation section and evaluated the validation in other studies. By selecting soft threshold based on free topology fit (Zhang and Horvath, 2005) and cutting hierarchical tree using dynamicTreeCut R package (Langfelder et al., 2008), 12 modules were detected in METABRIC by DiffCoEx. The gene–gene pairwise correlation distributions were calculated for METABRIC as well as the other four studies and the boxplots are shown in Figure 4. Most detected modules only showed moderate degree of validation.

3.3 Breast cancer studies (ILC versus IDC)
We finally applied our method to search for DCN between two breast cancer histological subtypes: ILC (invasive lobular carcinoma) and IDC (invasive ductal carcinoma). IDC and ILC are the two most common subtypes of breast cancers, representing 60-75% and 5-15% of all breast cancer cases, respectively (Guigu et al., 2014). Several studies have shown that they are two biological distinct diseases by comparing their genomic profiles, but the biological process driving for different subtypes are still largely unknown (Michaut et al., 2016). Identifying differential co-expression network between ILC and IDC can potentially unveil different biological mechanism and provide targets for precise treatment for ILC.

Using similar parameter settings, with the optimal weights and \( R = 10 \) repeats, at FDR \( \leq 0.3 \), 11 basic DC modules were detected as over-connected in IDC, and no modules were detected as over-connected in ILC. Pathway-enrichment-guided module assembly was performed for varying number of repeats with different initial seed modules. The results were quite consistent. We identified 4 supermodules engaged in 5 pathways, sharing at least 3 overlapping genes with enriched pathway. The top pathways associated with the assembled modules from 10 repeats were listed in Table 2(B) (see full list in Supplementary Table S3). Supplementary Figure S2 (A, B)
shows the visualizations of two supermodules enriched in protease inhibitor activity and proteinaceous extracellular matrix pathways. We also validated the densities and correlations of the basic modules ensembled in those two pathways in the validation sets (see Supplementary Fig. S2).

In the literature, alteration of extracellular matrix in tumor stroma has been shown relevant to metastatic potential (Oskarsson, 2013). Previous imaging analysis has further demonstrated different evolution of fibrillary collagen changes in ILC versus IDC throughout tumor progression (Burke et al., 2013).

4 Conclusion

In this study, we proposed a method, MetaDCN, to detect consensus differential co-expression (DC) networks across multiple studies with respect to certain phenotype of interest (e.g. case versus control or ER+ versus ER-). The method optimizes a target function to detect biologically meaningful DC modules. Since global optimization is computationally infeasible and unstable, we developed a simulated annealing algorithm to detect small (size 3–30) basic DC modules and assessed their false discovery rate. Through a pathway-guided module assembly algorithm, basic DC modules passing FDR threshold were merged into DC supermodules that were enriched in certain pathways to allow biological interpretation and hypothesis generation. The module assembly approach also allowed over- and under-connected basic DC modules to be simultaneously merged in a DC supermodule, representing possible alternative sub-pathway activation under different phenotypic conditions. Simulations and multiple transcriptomic studies is promising to identify biological meaningful and verifiable DC modules. MetaDCN meets the urgent need for this purpose and is expected to elucidate novel mechanisms in many disease investigations.

Funding

National Institutes of Health [R01CA190766 to L.Z., Z.H. and G.C.T.]; L.W. was supported by China Scholarship Council [201308110031] and National Nature Science Foundation of China [11126146]; C.Y.C. was supported to visit G.C.T.’s lab by Ministry of Science and Technology in Taiwan (103-2917-I-002-020); S.O. is a Susan G Komen Scholar, and supported by BCRF.

Conflict of Interest: none declared.

References


Fig. 4. Gene-gene pairwise correlation distribution of 12 modules detected by applying DiffCoEx on METABRIC. Solid lines separate training sets (S4) and testing sets (S1, S2, S3, S5) (Color version of this figure is available at Bioinformatics online.)