Gene expression

Statistical methods for gene set co-expression analysis

YounJeong Choi and Christina Kendzior

Department of Statistics and Department of Biostatistics and Medical Informatics,
University of Wisconsin - Madison 1300 University Avenue, Madison, WI 53706, USA

Received on April 2, 2009; revised on July 20, 2009; accepted on August 4, 2009
Advance Access publication August 18, 2009

ABSTRACT

Motivation: The power of a microarray experiment derives from the identification of genes differentially regulated across biological conditions. To date, differential regulation is most often taken to mean differential expression, and a number of useful methods for identifying differentially expressed (DE) genes or gene sets are available. However, such methods are not able to identify many relevant classes of differentially regulated genes. One important example concerns differentially co-expressed (DC) genes.

Results: We propose an approach, gene set co-expression analysis (GSCA), to identify DC gene sets. The GSCA approach provides a false discovery rate controlled list of interesting gene sets, does not require that genes be highly correlated in at least one biological condition and is readily applied to data from individual or multiple experiments, as we demonstrate using data from studies of lung cancer and diabetes.

Availability: The GSCA approach is implemented in R and available at www.biostat.wisc.edu/~kendzior/GSCA/

Contact: kendzior@biostat.wisc.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

A main goal of microarray experiments is to identify individual genes or gene sets differentially regulated across biological conditions. Most often, differential regulation is taken to mean differential expression; and a number of statistical methods for identifying differentially expressed (DE) genes or gene sets are now available (for reviews, see Allison et al., 2006; Barry et al., 2008; Ho et al., 2007; Newton et al., 2007). Although useful in thousands of studies, these methods are not able to identify many important classes of differentially regulated genes. One example concerns differentially co-expressed (DC) genes.

Two genes are DC if their correlation in one biological condition differs from that in another; and statistical methods for identifying DC gene pairs are available (Lai et al., 2004; Shedden and Taylor, 2005). Generally speaking, a DC gene group is defined similarly, as one in which the correlation structure among the group’s genes in one condition differs from that in another. However, the exact way in which one defines the gene group, specifies the correlation structure, and quantifies differences varies from study to study.

A number of investigators have proposed approaches that identify groups of genes where pairwise correlations are necessarily high in at least one biological condition (Brown et al., 2002; Choi et al., 2005; Ihmels et al., 2005; Kostka and Spang, 2004; Oldham et al., 2006; Watson, 2006). Most of the methods begin by identifying modules (Oldham et al., 2006), clusters (Brown et al., 2002; Choi et al., 2005; Ihmels et al., 2005; Watson, 2006) or cliques (Voy et al., 2006) within biological condition followed by a comparison of the condition-specific lists. Those modules, clusters or cliques identified in one condition but not another are of primary interest and often investigated further to determine if there is evidence of enrichment of biological function(s) potentially associated with the underlying mechanisms giving rise to the DC.

In this work, we propose a statistical approach to identify DC gene groups. Unlike previous work, our approach does not require that genes within a set are highly correlated in at least one biological condition; they may be, but differential regulation can manifest itself in significant but more subtle correlation shifts. Furthermore, the approach provides a false discovery rate (FDR) controlled list of interesting groups, and is readily applied to data from individual or multiple experiments. As detailed in Section 2, the approach requires that gene groups be defined a priori. We consider groups, referred to hereinafter as gene sets, specified by Gene Ontology (GO; The Gene Ontology Consortium, 2000) and the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000), noting that other annotations or study-specific biological knowledge could also be used. Pairwise co-expressions (correlations) are calculated for all gene pairs within a gene set; and a dispersion index is introduced to quantify the difference between the resulting gene set co-expression vectors. This gene set co-expression analysis (GSCA) is illustrated in the context of a single experiment in Section 2.1; applications to multiple experiments are provided in Section 2.2. Once DC gene sets are obtained, it is of interest to identify the specific genes within each gene set contributing most to the observed DC. A statistical test for identifying DC hub genes, or genes with unusually high contributions, is given in Section 2.3. A small simulation study and results from analyses of lung cancer and diabetes datasets are given in Sections 3 and 4, respectively. Section 5 concludes with a discussion of the advantages and disadvantages of the GSCA approach, and the similarities and differences to the well-known gene enrichment methods.
2 METHODS

The GSCA approach begins with a collection of gene sets. These can be
defined from GO, KEGG or some other biological knowledge. Of primary
interest is the identification of those gene sets significantly DC across
biological conditions.

2.1 Identification of DC gene sets within a single
experiment

Consider first a single two group microarray experiment. To assess the
degree of DC for a given gene set $c$ with $n_c$ genes, pairwise co-expressions
(correlations) are calculated for all $n_c(n_c-1)/2$ gene pairs, and a dispersion index
is applied to the co-expression vectors to quantify the extent of DC. A schematic
is given in Figure 1.

The dispersion index for a single study GSCA, $D_{c}$, is given by the
Euclidean distance, adjusted for the number of the gene set considered:

$$D_{c}(p_{c1}^{1}, p_{c2}^{2}) = \sqrt{\frac{1}{2n_{c}(n_{c}-1)/2} \sum_{i=1}^{n_{c}(n_{c}-1)/2} \rho_{c1}^{1} \rho_{c2}^{2}},$$

where $\rho_{c1}^{1}$ indexes gene pairs within the

gene set of size $n_c$, and $\rho_{c2}^{2}$ denotes the co-expression calculated for gene
pair $p$ within condition $T_k$, $k = 1, 2$. For a study with more than two conditions,
$D_{c}$ is averaged across study pairs.

To identify significant DC gene sets, samples are permuted across
conditions to simulate the null of equivalent correlation between conditions.
The GSCA approach shown in Figure 1 is applied to calculate a DC score
from the permuted dataset. This is repeated on $B = 1$ permuted datasets
to yield gene set-specific $P$-values. For example, for gene set $c$, the
permutation $P$-value is

$$P_{c} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}(\mathbf{A}^{b}, \mathbf{A}^{1}) \geq \mathbb{I}(\mathbf{A}^{1}, \mathbf{A}^{2})}{B},$$

where $\mathbf{A}^{1}$ and $\mathbf{A}^{2}$ denote samples derived from the $b$-th permuted dataset. An estimated
FW$E$ is obtained by converting the $P$-values to $q$-values (Storey and Tibshirani, 2003). In our simulations and case studies, we considered
Pearson’s correlation coefficients and $B = 10000$.

2.2 Identification of DC gene sets across multiple
experiments

The GSCA approach can combine evidence from multiple experiments
to identify DC gene sets. We refer to this as a meta-GSCA. As different
experiments use different microarray platforms that often contain different
sets of genes and gene identifiers, the problem of gene matching—identifying
the genes in common across studies—must be addressed prior to meta-
GSCA. Gene matching is generally done by specifying a gene identifier
common to all experiments, matching on those identifiers, and then removing
genes that are not represented across all experiments. In addition to gene
matching, it is also necessary to summarize transcript-level expression which
is often measured using multiple probes. Common methods include taking
the brightest probe (Mah et al., 2004; Subramanian et al., 2005), the most
variant probe (Dallas et al., 2005; Raghavan et al., 2007; Zhang et al., 2007)
or the average across the probes (Lee et al., 2008; Parmigiani et al., 2004;
Wang et al., 2007). We use the average taken at the log level.

Once a set of common genes is identified, gene sets are defined for the
common genes and meta-GSCA proceeds similarly to that above, with a few
important differences. First, in single study GSCA, it is of primary interest
to identify those gene sets with large differences in correlation between
conditions. This is also important in the meta-GSCA, but equally important in
preservation of the difference across studies. In other words, for a meta-
GSCA combining two studies $S_1$ and $S_2$, we would like to identify DC
gene sets for which the difference in co-expressions within $S_1$, $(\rho_{S_1}^{1} - \rho_{S_1}^{2})$, is close to the co-expression differences in $S_2$, $(\rho_{S_2}^{1} - \rho_{S_2}^{2})$. We use a test
statistic similar to (1), where $\tilde{\rho}_{S_1}^{1}, \tilde{\rho}_{S_1}^{2}$ is replaced by $\tilde{\rho}_{S_1}^{K}, \tilde{\rho}_{S_1}^{K}$ for the two-study case:

$$D_{a}(p_{a1}^{1}, p_{a2}^{2}) = \frac{1}{P_{c}} \sum_{c=1}^{P_{c}} \mathbb{I}(\mathbf{A}^{K}, \mathbf{A}^{c}) \geq \mathbb{I}(\mathbf{A}^{c}, \mathbf{A}^{2}),$$

where $\tilde{\rho}_{S_1}^{K}, \tilde{\rho}_{S_1}^{K}$ represents the sign of the correlation
for gene pair $p$ in condition $T_k$, $k = 1, 2$. For studies with more than two conditions, $D_{a}$ is averaged across study pairs.

Unlike the single study GSCA, the gene sets that are most interesting in
the meta-GSCA are those with unusually small values of the statistic given
by (2), as these are the sets that are most highly preserved across studies.
Note that gene sets containing many uncorrelated genes could appear to be
highly preserved, even if they are not, if $P_{c}$ is used as in (1). This is because
observed correlations for such sets would most often be near zero and, as a
result, the differences in correlations between studies would be necessarily
small. By considering $\tilde{\rho}_{S_1}$ instead of $\rho_{S_1}$, Equation (2) helps to ensure that
such sets are not identified. As in the single study GSCA, permutations are
used to calibrate the statistic given by (2). However, in the meta-GSCA
case, the null is that $(\tilde{\rho}_{S_1}^{1}, \tilde{\rho}_{S_1}^{2})$ differs from $(\rho_{S_1}^{1}, \rho_{S_1}^{2})$. Permuting samples
within each study across conditions results in preservation of $(\rho_{S_1}^{1}, \rho_{S_1}^{2})$
across studies since the values of $(\rho_{S_1}^{1}, \rho_{S_1}^{2})$ within each study $S_i$ will be near zero.
In other words, permuting samples across conditions as in a single study
GSCA breaks the DC structure which simulates the alternative, not the null.
Instead, we permute gene pairs within study across gene sets keeping the
gene set sizes fixed (see Supplementary Fig. S1). This preserves the overall
amount of DC, but breaks the relationship among gene pairs across studies.

2.3 Identification of DC hub genes

Given DC gene sets obtained from a single study or meta-GSCA, it is often of
interest to identify specific genes within the gene sets that contribute most to the
detected DC. Consider a gene $g$ within gene set $c$. For $K$ studies, a simple
ordering ranks $g$ according to the average DC, $a_{DC} = \sum_{i=1}^{K} \sum_{j=1}^{p} \mathbb{I}(\rho_{j}^{i}, \rho_{j}^{i})/n_c$, where $i$ indexes study and $j$ indexes the $n_c-1$ gene pairs containing $g$. A complementary approach that is less sensitive to outliers considers the
number of gene pairs containing $g$ with co-expression differences that
exceed the median of all co-expressions in $c$ (co-expressions are averaged
across studies in the case of multiple studies). In other words, we consider
$$s = \sum_{i=1}^{K} \sum_{j=1}^{p} \mathbb{I}(\rho_{j}^{i} \neq \rho_{j}^{i}, \text{median}(\rho_{j}^{i}))$$
where $p$ indexes the $P_{c}$ gene pairs within gene set $c$, as in (1). A hypergeometric distribution
can be used to calculate the probability that $j$ of $n_c-1$ gene pairs chosen from $P_{c}$ gene pairs exceed the median absolute correlation of the $P_{c}$ pairs.
Gene-specific $P$-values are obtained from the hypergeometric test given in
(3) and adjusted using a Bonferroni correction for $n_c$, the total number of
Specifically, we consider the two most common DE meta-analysis methods, to multiple studies, meta-GSCA is compared with an alternative approach. (Newton and can therefore be applied when replicate measurements are not available approach that shares information across genes (or in this case gene pairs) replicate measurements for a pair of genes determine a single co-expression gene pair has a single co-expression value within a given condition (i.e. using a hypergeometric test as in Falcon and Gentleman (2007). As each analysis (GSEA), and Barry et al. (2005), who proposed a framework for the significance analysis of function and expression (SAFE). The single study GSCA approach is most similar to GSEA (or SAFE) in that a single statistic is calculated for each gene set and calibrated via permutations across samples. The results of single study GSCA are therefore compared with GSEA in Section 4.1. We also compare the single study GSCA results to those obtained by testing for enrichment among DC gene pairs. In short, we evaluate condition-specific co-expression for all gene sets for which there are more DE genes than expected by chance. Evaluation proceeds through a hypergeometric test, or something similar (Falcon and Gentleman, 2007). A second enrichment approach considers all genes (not just DE genes) and identifies gene sets for which a set-level statistic looks unusual compared with the same statistic evaluated following label permutations. Details of this type of approach are given in Subramanian et al. (2005), who proposed the gene set enrichment analysis (GSEA), and Barry et al. (2005), who proposed a framework for the significance analysis of function and expression (SAFE). The single study GSCA approach is most similar to GSEA (or SAFE) in that a single

\begin{equation}
\sum_{j=1}^{nc} \left( \mathbf{P}_1 \mathbf{C}_1^{(j)} \mathbf{P}_2 \mathbf{C}_2^{(j)} \right)
\end{equation}

where \( \mathbf{P}_j = \{ \mathbf{P}_j \} \). When \( \mathbf{P}_j \) is odd, \( \mathbf{P}_j = -1/2 \) and \( \mathbf{P}_j = 1/2 \) are used for the left and right \( \mathbf{P}_j \), respectively, in the numerator. Genes with significant Bonferroni corrected hypergeometric P-values are referred to as hub genes.

### 4.1 Lung cancer

We illustrate the GSCA approach using the three lung cancer microarray datasets considered in Parmigiani et al. (2004) and Subramanian et al. (2005) and described in detail in Garber et al. (2001), Bhattacharjee et al. (2001) and Beer et al. (2002). Briefly, the three studies referred to here as the Stanford (Garber et al., 2001), Harvard (Bhattacharjee et al., 2001) and Michigan studies (Beer et al., 2002), were aimed at characterizing lung tumor gene expression profiles relative to that of normal lung tissue. The Stanford and Harvard studies include many subtypes of lung cancer, while the Michigan study focuses on lung adenocarcinomas, a tumor subtype included in the other two studies. The Harvard, Michigan and Stanford studies contain 17, 10, 5 normals and 139, 86, 41 tumor samples, respectively.

We considered the Entrez Gene ID, Unigene ID and Gene Symbol for gene matching. The Entrez Gene IDs were used as this ID gave the biggest inter-study gene coverage overlap for the lung cancer data. Following gene matching, the 3924 genes that appeared in all three studies were annotated into 3649 gene sets including 3471 GO categories and 178 KEGG pathways of at least size 3. We note that GSCA conducted within each study would not require any gene matching; however, gene matching was done here prior to all analyses to facilitate comparison of the GSCA results with the meta-GSCA results that follow.

The GSCA approach was applied to each of the three studies in isolation. Table 1 shows the total number of DC gene sets identified within each study at varying levels of FDR. Given the

<table>
<thead>
<tr>
<th>FDR (%)</th>
<th>Harvard</th>
<th>Michigan</th>
<th>Stanford</th>
<th>H&amp;M</th>
<th>All three</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>312</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1663</td>
<td>1582</td>
<td>0</td>
<td>947</td>
<td>0</td>
</tr>
</tbody>
</table>
Gene set co-expression analysis

![Figure 2. GO:0006955 immune response. The upper (lower) panel shows the 30 (6) genes most DC between cancer and normal. Edges represent co-expressions ranging from −1 (blue) to 1 (red). Nodes identified as DE by Rhodes et al. (2002) or Choi et al. (2003) are shaded.](image)

![Figure 3. Study specific GSCA q-values are shown for the 3649 gene sets (upper panels show varying angles of the 3D plot). Red, blue and light blue values correspond to gene sets for which the meta-GSCA q-values are q < 0.1, 0.1 ≤ q < 0.3 and q ≥ 0.3, respectively.](image)

![Figure 4. The average DC between two biological conditions is shown for 50 of the 211 genes in GO:0006955. The genes are ordered by P-values obtained from a hub-gene test (see Section 2.3). Red bars highlight genes with Bonferroni corrected P < 0.05.](image)


dispersion index specified in (1), the identification of a given set as DC could be due to a small number of gene pairs showing large differences in correlation between conditions, to many gene pairs showing moderate difference, or both. As a result it is useful to further investigate the identified gene sets to gain insight into specific sources of DC.

Consider a particular gene set, the immune response gene set GO:0006955, which was identified as DC at FDR 4.2% using the Harvard study data. To focus on a subset of the 211 genes in this set, genes were rank ordered by P-values derived from the hypergeometric test described in (3). The 30 genes with smallest P-values are shown in the upper panel of Figure 2. A striking feature concerns the presence of relatively stronger co-expressions in the normal condition. A closer look at a subset of the network (lower panel) highlights a few specific differences. For example, the co-expression between CFP and SKAP1 increases in cancer compared with normal; the opposite holds for CFP and RBM4. CFP, complement factor properdin, is a member of the properdin family which is known to play an important role in the immune system (Ivanovska et al., 2008; Stover et al., 2008) and has been associated with numerous types of cancer (Rottino et al., 2006).

Similar results are observed in the Michigan and Stanford studies (see Supplementary Fig. S3), although this gene set did not reach the same level of statistical significance. Using the Michigan data, the gene set is identified as DC at FDR 8.3%. With the Stanford data, the estimated FDR for this gene set is 49%, which is clearly quite high. However, we note that 0.49 was the smallest q-value observed in the Stanford DC analysis, largely due to the relatively small sample size.

When the data are combined in a meta-GSCA, the immune response gene set GO:0006955 as well as a number of others (48 at 5% FDR, shown in Supplementary Table S1) is identified as significantly DC. As in the case of GO:0006955, the sets identified in the meta-GSCA are largely those showing moderate, but not necessarily statistically significant evidence of DC within each study. This is shown in Figure 3, where the study-specific GSCA q-values are plotted for each study, and color-coded according to the meta-GSCA q-values. As shown, most of the gene sets identified as statistically significant in the meta-GSCA are those having relatively small (although not necessarily significant) study-specific q-values.

Figure 4 displays the gene-specific average DCs for 50 of the 211 genes in GO category GO:0006955, rank ordered by the hub-gene test described in Section 2.3. The most significant hub gene identified in this set is LCK, lymphocyte-specific protein tyrosine kinase, a much studied gene that is associated with lung and other kinds of cancer (Harashima et al., 2001; Imai et al., 2001; Krystal et al., 1998; Naito et al., 2007). Slow decay of the average DCs as shown suggests that there is not a single gene, or a few genes, driving the DC call for this dataset, but rather many genes showing a similar amount of DC overall. Investigation of such plots can be useful when identifying DC gene sets for which there are a few genes giving rise to a majority of the observed DC.

A similar calculation was carried out for each of the 48 gene sets identified in the meta-GSCA. The top 10 hub genes (10 genes with smallest hub-gene test P-values) were recorded for each set and the 12 genes showing up at least five times in the top 10 across the 48 sets are shown in Table 2. There we give the gene name and the number of sets (out of 48) for which that gene is in the top 10 hub-gene list. As genes present in many gene sets are favored for over
Common hub genes identified in 48 gene sets

Table 2. Common hub genes identified in 48 gene sets

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Top 10</th>
<th>GS</th>
<th>O-GS</th>
<th>O-A-GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MXI1</td>
<td>10</td>
<td>11</td>
<td>92</td>
<td>314</td>
</tr>
<tr>
<td>FADS1</td>
<td>10</td>
<td>10</td>
<td>113</td>
<td>427</td>
</tr>
<tr>
<td>RBM4</td>
<td>9</td>
<td>9</td>
<td>165</td>
<td>592</td>
</tr>
<tr>
<td>TGFB1</td>
<td>8</td>
<td>30</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BMP7</td>
<td>8</td>
<td>9</td>
<td>165</td>
<td>592</td>
</tr>
<tr>
<td>MICA</td>
<td>7</td>
<td>10</td>
<td>113</td>
<td>427</td>
</tr>
<tr>
<td>CFP</td>
<td>7</td>
<td>10</td>
<td>113</td>
<td>427</td>
</tr>
<tr>
<td>FEZ1</td>
<td>7</td>
<td>7</td>
<td>258</td>
<td>1064</td>
</tr>
<tr>
<td>CLOCK</td>
<td>7</td>
<td>7</td>
<td>258</td>
<td>1064</td>
</tr>
</tbody>
</table>

Shown are the gene name, the number of times (out of 48) the gene appears in the top 10 hub-gene list (Top 10), the number of gene sets (out of 48) containing that gene (GS), the number of other genes that appear in GS gene sets (O-GS), and the number of other genes that appear in at least GS gene sets (O-A-GS).

representation, we also report the number of sets (out of 48) that the gene appears in, the number of genes that appear in that many sets and the number of genes that appear in at least that many sets. For example, MXI1 appears in the top 10 genes in 10 of 48 gene sets. It is present in 11 of the 48 gene sets; 92 other genes are present in exactly 11 of the 48 gene sets; and 314 genes are present in 11 or more of the 48 gene sets. MXI1 is a well-known tumor suppressor gene (Ariyanayagam-Baksh et al., 2003; Eagle et al., 1995; Kim et al., 1998; Petersen et al., 1998), and has recently been studied with respect to its interactions with other genes (Corn and El-Deiry, 2007; Dang et al., 2008; Dooley et al., 1995; Tsao et al., 2008). A number of other interesting genes made the list, including TGFB1 which has been associated with lung cancer risk (Park et al., 2006), and BMP7, a gene recently identified as a potential therapeutic target for breast cancer (Yan and Chen, 2007) and metastatic bone disease (Buji et al., 2007).

We note that the results discussed here are largely distinct from those obtained from traditional enrichment methods (for details on the enrichment methods employed here, see Section 2.4). GSEA applied to the Harvard data found no gene sets to be enriched for DE genes at FDR 5% or 10% FDR; the smallest $q$-value from the GSEA analysis was 0.18. The upper left panel of Supplementary Figure S4 suggests that most of the gene sets show little DE between tumor and normal. That is not the case for DC (upper right panel of Fig. S4). Tests for enrichment among DC gene pairs gave analogous results with only eight gene sets identified at 5% FDR, compared with 312 gene sets identified by GSCA (Table 1). Two of the eight are represented in the 312; all eight are represented in the 1663 sets identified by GSCA at FDR 10%. A similar finding was observed in the meta-analysis. Rhodes et al. (2002) and Choi et al. (2003) identified 111 and 1534 of the 3924 genes to be DE at FDR 5%, with each of the 111 genes contained in 1534. The GO category GO:0006955 highlighted in Figure 2 contained five genes identified as DE using the method of Rhodes et al. (2002); the method of Choi et al. (2003) identified 86 DE genes. Given these totals, GO:0006955 was not found to be enriched for DE genes using the hypergeometric approach described in Falcon and Gentleman (2007) at FDR 5%. Indeed, neither the DE list derived from Rhodes et al. (2002) nor Choi et al. (2003) showed enrichment for any of the 3649 gene sets at FDR 5%; and, as shown in Supplementary Figure S5, the discrepancy between the meta-GSCA and enrichment tests is not due to the FDR threshold.

4.2 Diabetes

We performed a second meta-analysis with diabetes data obtained by searching the public repository NCBI GEO (Gene Expression Omnibus) and DGAP (Diabetes Genome Anatomy Project). As of February 29, 2008, NCBI GEO returned 79 GEO Series (GSE) for the search term ‘diabetes’. After removing series only peripherally related with diabetes, series without Entrez Gene ID annotation, series with fewer than three biological replicates and series without raw data files uploaded, 16 datasets from human, mouse and rat remained eligible for analysis. DGAP provided an additional six datasets for which the diabetic and normal conditions were clearly described. A brief summary of the 22 datasets is given in Supplementary Table S2. After gene matching by NCBI HomoloGene build 61, the 22 datasets represent 2349 common genes and 2253 common gene sets defined by GO and KEGG of size at least 3. We further reduced the collection to include eight experiments with sample size larger than 5, based on the shape of the distribution of correlation coefficients obtained from simulations (Supplementary Fig. S6). The final eight sets are marked in Supplementary Table S2. Meta-GSCA on the eight datasets identified 47 gene sets significantly preserved across studies at 5% FDR (sets are shown in Supplementary Table S3). The approach again identifies what are likely biologically meaningful gene sets. For example, the KEGG pathway for mitogen-activated protein kinase (MAPK) signaling was identified. This pathway is known to play a key role in both types 1 and II diabetes (Evans et al., 2003; Wellen and Hotamisligil, 2005). NR4A1, nuclear receptor subfamily 4, group A, member 1 in particular, included in this and many other significant sets, has been reported to be a regulator of hepatic glucose metabolism (Pei et al., 2006). PDK4, pyruvate dehydrogenase kinase, isozyme 4a, well known to be associated with diabetes (Cadoz et al., 2008; Kim et al., 2006), is another gene that also appears in many significant gene sets.

A particularly interesting gene set among the 47 is GO:0007169. Figure 5 shows a subset of six important genes selected from GO:0007169 as done for the lung cancer results shown in the lower panel of Figure 2. The gene set contains ERBB4, a gene known to be involved in pancreatic islet cell development (Huotari et al., 2002; Kritzik et al., 2000; Miettinen et al., 2000), which our group
has recently shown to be predictive of type 2 diabetes (Keller et al., 2008). In the normal condition, ERBB4 is non-negatively correlated with CRYAB and PRKCA in seven of the eight studies; the correlations are largely negative in the diabetic condition.

5 DISCUSSION

Statistical methods for identifying DE genes were among the first developed for microarray data, with methods for detecting enrichment following soon thereafter. Many methods to detect enrichment (e.g. hypergeometric test) involve identifying DE genes and then gene sets for which there are more DE genes than expected by chance; others (e.g. GSEA; SAFE) consider all genes and calibrate set-level statistics via label permutations. The single study GSCA approach proposed here is most similar to GSEA (or SAFE) in that a single statistic is calculated for each gene set and calibrated via permutations across samples. A major difference is that unlike GSEA (or SAFE), the GSCA statistic evaluates pairwise co-expression, as opposed to gene-specific expression, across a gene set. The gene sets identified as a result are those showing distinct correlation profiles across conditions, which may or may not be related to differences in average expression. As a result, the GSCA provides complementary information to traditional GSEA approaches and should be done in addition to, not in lieu of, a GSEA.

Implementation of GSCA requires that a number of decisions be made. The most important ones concern choosing measures of correlation and dispersion. The results reported here were obtained using Pearson’s correlation coefficients, although we note that a number of other measures could prove useful. For the lung cancer data considered, GSCA using Spearman’s correlation coefficients resulted in an increased number of gene sets identified (specific results not shown). For example, 1055 gene sets were identified for the Harvard study data (268 agree with the 312 identified using Pearson’s coefficients). A consideration of transformed correlation coefficients as well as alternative forms of the test statistic could also prove useful in the context of GSCA, particularly if the identification of specific correlation structures is of interest. Further simulation studies are required to provide general guidelines on the advantages and disadvantages of various implementations.

As with many dispersion indices one could consider, the ones proposed in Equations (1) and (2) can achieve significance for gene sets with a few highly DC gene pairs as well as those showing moderate evidence of DC across many pairs. As a result, it is interesting, informative and important to closely investigate the DC gene sets identified. The graphical summaries presented here can provide some insight into the gene pairs most DC across conditions.

In summary, the GSCA approach provides an FDR controlled list of gene sets DC between two or more biological conditions. Unlike previous work (Brown et al., 2002; Choi et al., 2005; Ihmels et al., 2005; Kostka and Spa%ng, 2004; Oldham et al., 2006; Watson, 2006), the GSCA approach does not require that groups of genes be highly correlated in at least one biological condition. This feature is an important one, as gene pairs in known regulatory pathways often show relatively low correlations overall (see Supplementary Fig. S7). A second important feature of GSCA is that multiple studies are naturally accommodated. In particular, the consideration of co-expression facilitates combining data across potentially different platforms since the measurements for analysis (i.e. co-expressions) are necessarily on the same scale. Finally, the computational simplicity of the GSCA lends itself to larger problems. We have here considered two group analyses, but are currently working on applications to gene set mapping, where groups are defined by genotypes at a genetic marker.

6 CONCLUSIONS

The GSCA approach provides an FDR controlled list of gene sets DC between two or more biological conditions. It does not require that groups of genes be highly correlated in at least one biological condition and can be applied within a single study or across multiple studies. It should prove useful as a complement to traditional enrichment methods.

ACKNOWLEDGEMENTS

We thank Michael Newton for comments that greatly improved the manuscript.

Funding: National Institute of Diabetes and Digestive Kidney Diseases (grant 58037); National Institute of General Medical Sciences (grant 76274).

Conflict of Interest: none declared.

REFERENCES

Naito, M.
Miettinen, P.J.
et al
Lee, H.
et al
Lai, Y.
Gentleman, R.
et al
Kim, Y.I.
Gentleman, R.


Huotari, M.-A.
et al


