Analysis and correction of crosstalk effects in pathway analysis

Michele Donato,1 Zhonghui Xu,2 Alin Tomoiağa,3 James G. Granneman,4 Robert G. MacKenzie,4 Riyue Bao,5 Nandor Gabor Than,6 Peter H. Westfall,3 Roberto Romero,2 and Sorin Draghici1,7,8

1Computer Science Department, Wayne State University, Detroit, Michigan 48084, USA; 2Perinatology Research Branch, NICHD/NIH, School of Medicine, Wayne State University, Detroit, Michigan 48201, USA; 3Center for Advanced Analytics and Business Intelligence, Texas Tech University, Lubbock, Texas 79409, USA; 4Center for Integrative Metabolic and Endocrine Research, Wayne State University, Detroit, Michigan 48084, USA; 5Department of Biological Sciences, Wayne State University, Detroit, Michigan 48084, USA; 6Department of Obstetrics and Gynecology, School of Medicine, Wayne State University, Detroit, Michigan 48201, USA; 7Department of Clinical and Translational Science, Wayne State University, Detroit, Michigan 48084, USA

Identifying the pathways that are significantly impacted in a given condition is a crucial step in understanding the underlying biological phenomena. All approaches currently available for this purpose calculate a P-value that aims to quantify the significance of the involvement of each pathway in the given phenotype. These P-values were previously thought to be independent. Here we show that this is not the case, and that many pathways can considerably affect each other’s P-values through a “crosstalk” phenomenon. Although it is intuitive that various pathways could influence each other, the presence and extent of this phenomenon have not been rigorously studied and, most importantly, there is no currently available technique able to quantify the amount of such crosstalk. Here, we show that all three major categories of pathway analysis methods (enrichment analysis, functional class scoring, and topology-based methods) are severely influenced by crosstalk phenomena. Using real pathways and data, we show that in some cases pathways with significant P-values are not biologically meaningful, and that some biologically meaningful pathways with nonsignificant P-values become statistically significant when the crosstalk effects of other pathways are removed. We describe a technique able to detect, quantify, and correct crosstalk effects, as well as identify independent functional modules. We assessed this novel approach on data from four experiments involving three phenotypes and two species. This method is expected to allow a better understanding of individual experiment results, as well as a more refined definition of the existing signaling pathways for specific phenotypes.

[Supplemental material is available for this article.]
**Crosstalk effects**

In order to demonstrate the existence and assess the extent of crosstalk effects, we conducted a systematic exploration of this phenomenon. We constructed a reference set of genes from all genes present in at least one KEGG signaling pathway (2963 genes at the time). Then, each pathway was used as a “bait,” choosing from it a number of genes that would make it significant at a chosen significance level ($\alpha = 0.01$ after Bonferroni correction). Other random genes were selected up to a constant number ($n = 100$) of “differentially expressed” (DE) genes (see Supplemental Material for details). Under these circumstances, the research hypothesis is true for the bait, while the null hypothesis is true for all other pathways. We repeated this selection 1000 times for each pathway $P_i$ and each time, we computed the Fisher’s exact test $P$-value (Tavazoie et al. 1999), SPIA (impact analysis) (Tarca et al. 2009), and GSEA (Subramanian et al. 2005) $P$-values for all pathways from the KEGG database (Kanehisa et al. 2004). With these results, we constructed the empirical distributions of the false discovery rate (FDR)-corrected $P$-values corresponding to each prey $P_j$. The distributions of the $P$-values for all three methods are severely skewed toward zero, showing that all methods produce a significant number of false positives due to crosstalk effects (Supplemental Fig. S1). We hypothesized that crosstalk is mostly due to the common genes between pathways. If this were true, we would expect to see a strong coupling between pairs of pathways with many genes in common and a weak coupling between pathways that do not share any genes. In order to test this hypothesis, we calculated the Jaccard similarity index (Jaccard 1910) for each pair of pathways, as well as the Pearson correlation between the Jaccard index and the $P$-values of the preys (Figs. 1, 2). The data shows a very strong correlation (Pearson correlation index of 0.87 for Fisher’s exact test, 0.62 for GSEA, and 0.83 for SPIA), which confirms our hypothesis that the crosstalk can be explained by the presence of genes that are involved in more than one pathway.

**Results**

**Crosstalk analysis and correction**

The method we propose for correcting for crosstalk effects takes as input a set of reference pathways and a list of genes that are DE in the given condition. The crosstalk analysis is composed of three steps. The first step is the computation of a crosstalk matrix, i.e., the analysis of pairwise crosstalk interactions between pathways in the given condition, resulting in a heat map showing the $P$-value of each pathway when the genes from each other pathway are removed. This matrix allows the visualization of various cases of crosstalk effects and provides information on the extent of crosstalk effects in the condition analyzed. The second step is the module detection that allows the identification of subpathways that appear to be involved in the specific condition, studied independently of the pathway they belong to. The final step is the maximum impact estimation, where the biological impact of each gene is assigned to only one of the pathways to which the gene belongs, for those genes that belong to more than one pathway. The result of the last two steps is essentially a new set of pathways that can include (1) original pathways as found in the literature, (2) novel functional modules that might be relevant to the given condition, and (3) pathways modified by the removal of a such a functional module. The new set of pathways is not affected by crosstalk effects, and new $P$-values are computed for each pathway, building a ranked list from which the false positives due to crosstalk have been removed (see Methods).

**Fat remodeling in mice**

We include here the results of the crosstalk analysis in an experiment investigating cellular and metabolic plasticity of white fat tissue (WAT), where the classical overrepresentation analysis (ORA) produced a number of false positives, and failed to rank...
highly pathways that were known to be involved in the given
condition. In this experiment, the chronic activation of WAT beta-
adrenergic receptors by certain physiological and pharmacological
conditions transforms the tissue into one resembling brown fat,
a thermogenic organ (Granneman et al. 2005; Li et al. 2005;
Mottillo et al. 2007). The data set was obtained from a microarray
analysis of white fat from mice treated with low dose (0.75 mmol/h)
CL 316,243 (CL) for 0, 3, and 7 d. Here we discuss only the list of DE
genes coming from the comparison between expression levels of
genomes at days 3 and 0 (for the comparison between days 7 and 0, see
Supplemental Material).

The top 20 pathways ranked by ORA and their associated
FDR-corrected \( P \)-values are shown in Figure 3A. In this figure,
pathways highlighted in red represent pathways not related with
the phenomenon in analysis, while pathways highlighted in green
are those for which we know, with reasonable confidence, that
they are involved in the given phenomenon. The white back-
ground indicates pathways for which we do not have conclusive
information on their involvement (or lack of) with the phenom-
enon in analysis. The three most significant pathways in the
comparison between days 3 and 0 were Parkinson’s, Alzheimer’s,
and Huntington’s diseases. The fourth pathway in the ranked list is
Leishmaniasis. The first three pathways describe degenerative dis-
eases of the central nervous system that have no connection to fat
remodeling. Leishmaniasis describes the signaling involved in
a disease spread by the bite of certain species of sand flies. Clearly,
this pathway is also unlikely to give insights about the fat remod-
eling phenomenon. While other pathways, such as Phagosome

**Figure 3.** The results of the ORA analysis in the fat remodeling experiment for the comparison between days 3 and 0, before (A) and after (B) correction for crosstalk effects. All \( P \)-values are FDR corrected. The lines show the significance thresholds: (blue) 0.01, (yellow) 0.05. Pathways highlighted in red represent pathways not related with the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of) with the phenomenon in analysis. (A) The top 20 pathways resulting from classical ORA before correction for crosstalk. The top four pathways are not related to fat remodeling. (B) The top 20 pathways after correction for crosstalk. Pathways ranked 1, 3, and 5 are modules that are functioning independently of the rest of their pathways in this particular condition. Starred pathways are pathways edited by removing such modules. Note the lack of any obvious false positive above the significance threshold(s).
(Newman et al. 1982), PPAR Signaling (Granneman et al. 2005), and Cell cycle (Lee et al. 2012), are definitely more related to the phenomenon of fat remodeling, their presence in the middle of a ranked list dominated by false positives (six false positives in the 10 pathways significant at 1%) illustrates how the crosstalk effects make the classical ORA unable to find the truly relevant pathways.

In order to analyze and eliminate the crosstalk effects, we computed the crosstalk matrix as described in the Methods section. The analysis of the matrix illustrates some interesting examples of crosstalk effects. Figure 4 represents a detail of the entire matrix. In this figure, the high significance of Parkinson’s (bright red in row 1, column 1) disappears when the crosstalk due to Alzheimer’s is eliminated (green in row 1, column 2). This indicates that Parkinson’s is a false positive, since its significance is due exclusively to genes from Alzheimer’s. Furthermore, the high significance of Alzheimer’s (bright red in row 2, column 2) also disappears when the crosstalk effect of Parkinson’s is eliminated (green in row 2, column 1). This means that Alzheimer’s significance is also due only to the genes in common with Parkinson’s. Essentially, the analysis tells us that the genes in common between the two pathways are activated independently of either pathway, which suggests that these genes constitute an independent functional module. The same phenomenon involves the Cardiac Muscle Contraction and Huntington’s disease pathways. The same independent functional module is responsible for the changes shown in areas marked with a in Figure 4.

An inspection of these genes and their signaling mechanisms reveals that this module is composed by genes present in mitochondria, organelles involved in all pathways above. The fact that this module is strongly activated in this fat remodeling experiment that is not related to any of the above conditions (Alzheimer’s, Parkinson’s, Huntington’s), suggests that this should be considered as an independent pathway, dedicated to mitochondrial activity. The second data set analyzed was obtained from a recent study that investigated the transcriptome of uterine cervical ripening in human pregnancy before the onset of labor at term (Hassan et al. 2009). The tissue analyzed is the human uterine cervix, the lower part of the uterus extending from the isthmus of the uterus into the vagina. It is mainly composed of smooth muscle and extracellular matrix, which consists of collagen, elastin, proteoglycans, and glycoproteins (Uldbjerg et al. 1983b; Leppert 1995). The uterine cervix has an essential function in the maintenance of pregnancy and also in parturition (Hassan et al. 2006, 2009, 2010). Cervical ripening is a critical component of the common terminal pathway of parturition, which includes the extensive remodeling of the cervix (Hassan et al. 2009). Disorders of cervical ripening can lead to premature or protracted cervical change, complicating term (e.g., protracted dilatation or arrest of dilatation) or preterm gestations (e.g., premature cervical dilatation in the second trimester) (Hassan et al. 2009). The state of cervical ripening has traditionally been assessed by clinical examination (Bishop score or its modifications) (Bishop 1964), which includes the digital examination of
the cervix for its consistency, dilatation, effacement, and position. This method has also been used to predict the likelihood that a patient would go into spontaneous labor. The goal of this experiment was to examine the relationship between human cervical ripening and the cervical transcription, aiming to improve our understanding of the biology of cervical ripening at term. This study included pregnant women who underwent elective C-section at term with an unripe (n = 11) or ripe cervix (n = 11). Cervical biopsies were obtained from these women transvaginally, from the anterior lip of the uterine cervix following C-section. Microarray analysis was performed on RNA isolated from these cervical tissue specimens using Affymetrix GeneChip Human Genome U133 Plus 2.0 Arrays (Hassan et al. 2009).

On this data set, we performed the comparison between gene expression levels from cervical tissues obtained from women with an unripe or ripe cervix using the classical ORA. The results are shown in Figure 7A. Pathways with a P-value smaller than 0.05 after FDR correction were Focal adhesion, ECM-receptor interaction, Amoebiasis, Cell adhesion molecules (CAMs), Small cell lung cancer, and Dilated cardiomyopathy.

There is plenty of experimental evidence that biological processes described by the pathways Focal Adhesion, ECM-Receptor Interaction, and Cell Adhesion Molecules are related to cervical ripening. The relation between these pathways and the phenomenon in analysis was revealed by studies on humans and animals showing the involvement of extracellular matrix metabolism and cell adhesion molecules in cervical ripening (Uldbjerg et al. 1983a,b; Leppert et al. 1986; Leppert 1995; Mahendroo et al. 1999). However, the pathway Amoebiasis describes the biological process of infection from a parasite that invades the intestinal epithelium. Amoeba infection involves the parasite attachment to the intestinal mucus layer, followed by disruption and death of host epithelial cells. This process is completely unrelated to the physiological condition of cervical ripening in term pregnancy. The same is true for the Small Cell Lung Cancer pathway. Clearly, the top ranked pathways include some describing complex phenomena that are unrelated to the studied condition. Also, the significant pathways known to be involved in the process of cervical ripening are somewhat general pathways describing cellular interactions.

The analysis of the crosstalk matrix (see Supplemental Fig. S6) shows that there is an independent functional module among the top three pathways in the ranking. This novel module includes the genes present in the interaction between the cellular transmembrane protein integrin and three important ECM components—collagen, laminin, and fibronectin. The KEGG pathways involved in the identification of this pathway are Focal adhesion, ECM-receptor interaction, and Amoebiasis. Henceforth, we will refer to this pathway as the Integrin-Mediated ECM Signaling (Fig. 8).

Very interestingly, the independent functional module found in this condition is, in fact, the exact same module found in the hormone treatment experiment described in the Supplemental Material. Interestingly, the KEGG pathways involved in the identification of this functional module are slightly different between the two phenotypes. While in this phenotype this module was found from the interaction of Focal adhesion, ECM-receptor interaction, and Amoebiasis, in the hormone treatment the last pathway is replaced by Pathways in Cancer. The fact that the same module was found to be activated and statistically significant in two different phenotypes, from the interaction of different sets of canonical pathways, further supports the idea that this module describes an independent mechanism and should therefore be considered as an independent pathway.

Further analysis of the crosstalk matrix shows that the Small Cell Lung Cancer loses significance when the crosstalk effects of the first three pathways are removed (bright green loss of significance in first three columns of row 5 in Supplemental Fig. S6). This allows us to conclude that it is a false positive in the classical ORA, with its ORA significance due exclusively to crosstalk effects.

The ranking of pathways with the P-values corrected for crosstalk by our analysis is shown in Figure 7B. The first pathway is Integrin-mediated ECM Signaling with an FDR-corrected P-value of $2.9 \times 10^{-13}$. Cell Adhesion Molecules is now the second in ranking, with an FDR-corrected P-value of 0.004. The false positives in the classical ORA results, Amoebiasis and Small Cell Lung Cancer, are not significant anymore. The biological significance of the pathway Dilated Cardiomyopathy may be linked to the fact that 10%–15% of the uterine cervix is constituted of smooth muscle, and cervical ripening involves alterations of this component. The last significant pathway at the 5% significance threshold is Leukocyte Transendothelial Migration. Although human and animal studies (Hassan et al. 2009) have shown that cervical ripening does not require activation of a typical inflammatory response and influx of inflammatory cells into the cervix, the significance of this pathway may reflect the beginning of later inflammatory events typical of parturition (Word et al. 2007; Timmons et al. 2009).

![Figure 5](https://www.genome.org/)

**Figure 5.** Mitochondrial activity pathway. This independent functional module is responsible for the incorrect identification of the pathways Parkinson’s disease, Alzheimer’s disease, Huntington’s disease, and Cardiac Muscle Contraction by the classical ORA.

Figure 6. Epididymal white adipose tissue of a control mouse (left) and a mouse treated with CL for 7 d (right). Treatment with CL for 7 d triggered massive mitochondrial biogenesis, demonstrating in vivo that indeed, the mitochondrial pathway is central in this experiment. White bar, 20 μm.
In addition to the data sets described above, we analyzed a data set coming from the comparison of expression levels at days 7 and 0 in the fat remodeling experiment, a data set investigating the effect of various types of hormones on the endometrium of healthy post-menopausal women who underwent hysterectomy (Hanifi-Moghaddam et al. 2007), as well as a data set requested by one of the reviewers, investigating the correlation between gene expression values “in subjects with Alzheimer’s disease” in subjects with Alzheimer’s disease with MiniMental Status Examination (MMSE) and cortical atrophy (Blalock et al. 2004). The latter data set was analyzed to identify significant pathways from both KEGG and Reactome databases that were found to be independently activated and statistically significant in two different conditions: hormone treatment of post-menopausal women and cervical ripening in normal pregnancies. Genes shown in red were found to be differentially expressed in the hormone treatment experiment.

This approach is a departure from the current paradigm that considers the pathways as static models, independent of the phenotype. In the view proposed here, various specific modules, or sub-pathways, can be dynamically linked to specific conditions. When such independent functional modules are identified in independent conditions, such as the integrin-mediated ECM signaling above, these modules could be considered as candidate new pathways.

### Methods

#### Pathway analysis in the presence of overlapping pathways: The crosstalk matrix

Our goal is to develop an approach that can detect and quantify the crosstalk between pathways. The main issue we are trying to address here is the fact that, in the presence of overlapping pathways (i.e., for all pathways databases available today), crosstalk phenomena increase the probability of false positives, i.e., increase the number of pathways reported as significant but that in reality are not interesting (borrowing terminology from Brad Efron, we call pathways that have lesser biological significance “not interesting” even though they might be statistically significant with a large enough sample size). To better understand the approach we are going to present, let us briefly review the classical Fisher’s exact test approach described above. Figure 9A represents the contingency table used for assessing the significance of a pathway $P_i$ by the classical overrepresentation approach. The table divides genes as either being in the pathway or not, versus being considered DE or not DE (NDE); $n_i$ represents the number of DE genes on $P_i$, while $n$ represents the total number of DE genes, and $m_i$ represents the number of NDE genes on $P_i$, while $m$ represents the total number of NDE genes. It follows that $n_i + m_i = |P_i|$ represents the number of
Detection, analysis, and correction of crosstalk

A comparison of the classical overrepresentation analysis (4) with the crosstalk matrix analysis proposed here (8). (A) The standard overrepresentation approach contingency table: $n_i + m$ and $n + m$ represent, respectively, the number of genes belonging to pathway $P_i$ and the total number of genes, $n$ and $n$ represent, respectively, the number of differentially expressed genes belonging to pathway $P_i$ and the total number of DE genes. (B) Contingency table for the overrepresentation approach, taking into account the overlap between pairs of pathways: $P_{ij}$ represents the set of elements in $P_i$ excluding the intersection with $P_j$; with the notations $n_{ij} + m_{ij}$ we represent the total number of genes that are in pathway $P_i$ but not in pathway $P_j$, and with $n_{ij}$ the number of DE genes that are in pathway $P_i$ but not in pathway $P_j$.

Table 1: Contingency table for the overrepresentation approach, taking into account the overlap between pairs of pathways.

<table>
<thead>
<tr>
<th></th>
<th>DE</th>
<th>NDE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_i$</td>
<td>$n_i$</td>
<td>$m_i$</td>
<td>$n_i + m_i$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
<td>$m$</td>
<td>$n + m$</td>
</tr>
<tr>
<td>$P_{ij}$</td>
<td>$n_{ij}$</td>
<td>$m_{ij}$</td>
<td>$n_{ij} + m_{ij}$</td>
</tr>
<tr>
<td>$P_j$</td>
<td>$n_j$</td>
<td>$m_j$</td>
<td>$n_j + m_j$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
<td>$m$</td>
<td>$n + m$</td>
</tr>
</tbody>
</table>

The reasoning behind the ORA is that if the number of DE genes on a pathway is much higher than expected by chance, then the pathway is likely to be biologically interesting. In order to take into account the effect of the overlap on the significance of the two pathways, we consider the effect of the removal of the overlapping part on the significance of the pathways. This is achieved as follows: let us consider two overlapping pathways $P_i$ and $P_j$ with the notation $P_{ij}$ we define the set of elements in $P_i$ excluding the intersection with $P_j$; with the notations $n_{ij} + m_{ij}$, $n_{ij}$, and $m_{ij}$ we represent the number of genes that are in pathway $P_i$ but not in pathway $P_j$, and with $n_{ij}$ the number of DE genes that are on pathway $P_i$ but not in pathway $P_j$. We then consider the contingency table shown in Figure 9B, whose bottom margin is identical to that of Figure 9A.

With this contingency table, we compute for every pair of pathways $[i, j]$ the P-value of $P_{ij}$. Since this computation yields a $k \times k$ matrix, where $k$ is the number of pathways, the results are most conveniently represented using a heat map of the negative log P-values. Each cell $(i, j)$ of this matrix characterizes the significance of pathway $P_i$ when we remove the effect of pathway $P_j$. The rows and the columns are ordered by the original $P$-values of the pathways, which are placed on the diagonal. We will refer to this matrix as the crosstalk matrix. This matrix is useful for identifying the effects of crosstalk among pathways.

An example of the crosstalk matrix can be found in Figure 10. We will refer to the part of the matrix above the horizontal significance threshold as the significance strip. The nonsignificance strip will be the part below the horizontal significance threshold. The significance quadrant will be the part of the significance strip to the left of the significance threshold. Using these terms, we can identify and discuss several interesting phenomena that are not captured by any of the existing pathway analysis methods.

A first interesting case is when a pathway $P_i$ is reported as significant by the classical analysis, but it loses its significance when the effect of another pathway $P_j$ is removed. This is represented in the crosstalk matrix, by a nonsignificant P-value (green square) in the significance strip. In this case, $P_i$ is unlikely to be biologically meaningful, since its significance is most likely due to a crosstalk from $P_j$.

A second interesting case is when a pathway $P_i$ that is not significant for the classical analysis becomes significant when the crosstalk effect of another pathway $P_j$ is removed. This is represented in the crosstalk matrix by a significant P-value (red square) in the nonsignificant strip. The meaning of this is that pathway $P_j$ was masking the significance of $P_i$, indicating that a phenomenon likely to be biologically meaningful is happening in the part of $P_i$ which is not in common with $P_j$.

A third and last interesting case is a symmetric (with respect to the diagonal) decrease in significance of pathways in the significance quadrant. This indicates the presence of an independent functional submodule, common to both $P_i$ and $P_j$, that is responsible for their significance. Note that the activity of this module is tightly related to the condition studied.

The maximum impact estimation: An expectation maximization technique for the assessment of the significance of signaling pathways in presence of crosstalk

The crosstalk matrix is a useful tool for the interpretation of the effect of crosstalk between pathways. However, the ultimate goal of the analysis of signaling pathways is to provide a meaningful ranking among pathways, as well as a P-value quantifying the likelihood that a certain pathway is involved in the phenomenon in analysis. Here, we developed a correction method for the ranking of pathways that takes into account the overlaps between pathways.

The main idea is that if there is no crosstalk, then there is no ambiguity in the ORA significance calculations. In such a case, if genes in a pathway are overrepresented, it cannot be a false positive caused by crosstalk. Our approach is therefore to infer an underlying pathway impact matrix where each gene contributes to one and only one pathway and hence is devoid of crosstalk, and then to perform the ORA analysis using that impact matrix. Since this underlying pathway impact matrix is not observed directly, it is inferred through likelihood-based methods, and estimated using the EM algorithm. The corrected ranking is computed using ORA analysis with the underlying pathway impact matrix, shown as follows:

Let us consider the DE indicator vector $Y$, representing the differential expression of genes, and the membership matrix $X$ describing the membership of each gene in each one of $k$ pathways $P_1 \ldots P_k$. The vector $Y$ is defined as follows:

$$Y_i = \begin{cases} 1 & \text{if } g_i \text{ is DE} \\ 0 & \text{if } g_i \text{ NDE} \end{cases}$$

and each cell $X_{ij}$ of the matrix $X$ is defined as follows:

$$X_{ij} = \begin{cases} 1 & \text{if } g_i \text{ belongs to } P_j \\ 0 & \text{if } g_i \text{ does not belong to } P_j \end{cases}$$

Figure 10. Example of a crosstalk matrix. On the diagonal, we find the classical overrepresentation analysis, ordered by P-value. The blue line represents the 0.01 significance level, while the black line represents the 0.05 significance level. The P-values in the matrix have been log-transformed (base 10 log), and the sign of the result has been inverted. The color of the cell represents the P-value: bright red for P-values close to zero, bright green for P-values close to 1. 

www.genome.org
The matrix $Y|X$ obtained by combining the vector $Y$ with the $X$ matrix is shown in the example in Figure 11.

In many analysis methods, the membership matrix $X$ is also interpreted as the impact matrix: if $X_{ij} = 1$, then gene $g_i$ impacts pathway $P_j$. In ORA, for example, each gene is considered to have the same full impact on all pathways the gene belongs to. CROSSTALK effects result from the fact that a gene can belong to more than one pathway, but in principle, it can potentially have a different biological impact on each such pathway. Our aim is to identify the pathway where the biological impact of such a shared gene is maximum. We do so by estimating the maximum impact pathway using an expectation maximization approach (see Supplemental Material).

### Identification of independent functional modules

The maximum impact estimation procedure alone is not able to identify overlapping modules responsible for the entire significance of other pathways, as in the situations represented by case 3 in the section describing the crosstalk matrix. In such cases, the overlap should be considered as a separate pathway that is more likely to be biologically meaningful in the condition under analysis. An additional step is needed in order to correctly deal with this situation. In this additional step, we extract certain significant overlaps from the list of pathways and include them in the list as independent functional modules. An independent functional module is a module for which there is evidence of an activity independent of the pathways it resides in, for the given condition. If an independent module is found in more than one, possibly unrelated, condition, this module is considered as a candidate novel pathway.

A module must satisfy certain conditions in order to be treated as an independent functional module. Let us assume that we are analyzing the overlap between the pathways $P_i$ and $P_j$; the first condition is that both pathways are significant (after FDR correction for multiple comparisons) at a certain threshold $\alpha$. The threshold $\alpha$ is the significance threshold chosen by the user. Typical values for this threshold are 0.01 and 0.05. This condition limits the search to the significance quadrant of the crosstalk matrix. The second condition is that the overlap $P_{ij}$ itself must be significant at $\alpha$ (after FDR correction). The third condition is that the subpathways obtained by removing the overlap from both original pathways, indicated by $P_{ij}$ and $P_{ij}$ must not be significant at $\alpha$ (after FDR correction). If we denote with $p(P)$ the $P$-value of a generic pathway $P$ then the conditions can be summarized as follows:

- $p(P_i) < \alpha$, $p(P_j) < \alpha$, $p(P_{ij}) \geq \alpha$

This pairwise procedure might yield modules that are similar one to each other, for example, in cases where a module is contained in three or more pathways. That could be solved with a three-way or $n$-way search, but we opted for another approach for limiting the number of new modules. Once all interesting pairwise modules are created, we test for similarity among modules. The index used for similarity is a modified Jaccard similarity index mJS defined as follows:

$$mJS = \frac{|M_1 \cap M_2|}{\min\{|M_1|,|M_2|\}},$$

where $M_1$ and $M_2$ are two modules obtained with the search criteria explained above. We merge any two modules whose similarity is greater than a certain threshold $st$. Once the modules are merged, the similarity among all the modules (including the newly created one) is computed again, and the merging procedure is applied again until there are no more modules that can be merged.

These newly created modules are removed from all pathways with which they overlap, and this list of modified pathways is used in the EM procedure. For the data sets analyzed in this work, we used an $st$ threshold of 0.25. The procedure used to select this threshold can be found in the Supplemental Material.

After applying the module discovery and the EM approach, the result is a modified membership matrix that can be used to perform the desired type of analysis. This matrix now includes three types of pathways: (1) original pathways as found in the literature; (2) novel functional modules that are impacted in the given condition independently from the pathways they belong to; and (3) the pathways from which such independent modules have been removed. If the same independent module is found in several conditions—in other words, if this module is active independently from its parent pathways in several different phenotypes—such a module should be considered a good candidate for a novel pathway.

Additional aspects of the crosstalk correction, such as a detailed description of the module detection step, false negative rate, and pathway size bias, can be found in the Methods section of the Supplemental Material (module detection step and false negative rate), and in Supplemental Figures S11 and S12 (pathway size bias).

### Acknowledgments

This work has been partially supported by the following grants: NIH RO1 RDK089167, R42 GM087013, and NSF DBI-0965741 (to S.D.), by the Robert J. Sokol Endowment in Systems Biology, and by the Perinatology Research Branch, Division of Intramural Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NIH, or any other of the funding agencies.

**Author contributions:** M.D. and S.D. identified the problem, conceived the study, and wrote the manuscript. M.D. and Z.X. developed the crosstalk correction based on the EM. S.D., M.D., P.H.W., and A.T. developed the crosstalk matrix analysis. J.G.G. and R.G.M. performed the fat remodeling experiment, helped interpret the results, and edited the fat remodeling experiment section in the manuscript. R.B. helped interpret the results and edited the hormone treatment experiment section in the manuscript.

---

**Figure 11.** Example of a DE/membership matrix. Column $Y$ represents the indicator of differential expression of the various genes ($1$ for the $n$ DE genes and $0$ for the $m$ NDE). Column $P_i$ represents the membership indicator for pathway $i$. Row $g_j$ describes gene $j$ in terms of its differential expression and its membership in the various pathways.
N.G.T. helped interpret the results and edited the cervical ripening section in the manuscript. P.H.W. edited the crosstalk matrix analysis and the crosstalk correction section in the manuscript. S.D., P.H.W., and R.R. supervised the study.

References


Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. 2003. Global analysis and the crosstalk correction section in the manuscript. P.H.W. edited the crosstalk matrix analysis and the crosstalk correction section in the manuscript.


Received December 12, 2012; accepted in revised form August 6, 2013.
Analysis and correction of crosstalk effects in pathway analysis

Michele Donato, Zhonghui Xu, Alin Tomoiaga, et al.

Genome Res. published online August 9, 2013
Access the most recent version at doi:10.1101/gr.153551.112

Supplemental Material
http://genome.cshlp.org/content/suppl/2013/09/17/gr.153551.112.DC1.html

Published online August 9, 2013 in advance of the print journal.

Creative Commons License
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

© 2013 Donato et al.; Published by Cold Spring Harbor Laboratory Press