Supplementary materials - analysis and correction of crosstalk effects in pathway analysis

Michele Donato[1], Zhonghui Xu[2], Alin Tomoiaga[3], James G. Granneman[4], Robert G. MacKenzie[4], Riyue Bao[5], Nandor Gabor Than[6], Peter H. Westfall[3], Roberto Romero[2], Sorin Draghici[1,7*]

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

July 30, 2013

* *To whom the correspondence should be addressed. E-mail: sorin@wayne.edu, Fax: +1-313-577-6679*

1 Crosstalk phenomena

We hypothesized that pathways can consistently affect each other's p-values in significant ways through crosstalk. Identifying such effects in any number of specific real experiments would constitute only anecdotal evidence since the true amount of crosstalk between two given pathways in any given condition is not known. In order to demonstrate the existence and assess the extent of crosstalk effects, we designed and conducted the following systematic exploration of this phenomenon. We first constructed a reference set of genes from the union of all genes present on at least one KEGG signaling pathway (2963 genes at the time). Then, for each pathway P_i, we ran experiments as follows. We first calculated the number n_i of DE genes that would make P_i significant at least at $\alpha = 0.01$ after a Bonferroni correction for multiple comparison. Henceforth, we will refer to this pathway as the "bait". We then used the reference set to pick n_i random genes from P_i and $100 - n_i$ genes that are not on P_i, and calculated the Fisher Exact Test significance of all other "prey" pathways, P_j. This essentially models a situation in which 100 genes are found to be DE, and these genes are such that the Fisher Exact Test will find the bait pathway P_i significant at 1% after the correction for multiple comparisons. Since the $100 - n_i$ genes that are not on P_i are randomly chosen among the reference set, no other pathway P_j should have more genes than expected by chance. 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 1,000 times for each pathway P_i, and each time we computed the Fisher Exact Test, SPIA (impact analysis), and GSEA p-values for all pathways in the set S including all the pathways from KEGG. With these results, we constructed the distributions of the FDR-corrected p-values corresponding to each prey P_j. Under the null hypothesis, the p-values are expected to follow an uniform distribution, and to be independent between different pathways. In fact, the distributions of the p-values (see Fig. S1) are significantly different from the uniform distribution (Kolmogorov-Smirnov goodness of fit p-values of the order of 10^{-16} in all cases). The distributions for all three methods are severely skewed towards zero, showing that all methods produce a large number of false positives.

Furthermore, we observed much stronger crosstalk effects for specific pathway pairs (i, j): every time one of them is used as a bait, the p-value of the other one is pulled to values much lower than expected by chance, many times well below the significance threshold. All crosstalk effects can be represented in a crosstalk matrix (left panel in Fig. 1). In this matrix, the elements $[i, j]$ represent the mean of the distribution of p-values for 1,000 random trials using pathway i as bait and pathway j as prey. This matrix is not symmetrical since the influence of pathway i on pathway j can be different from the influence of pathway j on i. The matrix shows strong crosstalk between several pathways (e.g. row 3 and columns 57 through 70).

We hypothesized that this crosstalk is due mostly to the genes that are in common between pathways. If this were true, we would expect to see a strong coupling between pairs of pathways that have 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 between all pairs of signaling pathways from KEGG. The Jaccard index is defined as $\frac{|P_i \bigcap P_j|}{|P_i \bigcup P_j|}$, and characterizes the overlap between two sets, relatively to the size of their union. Pathways that share many genes will have a large Jaccard index. The right panel in Fig. 1 in the main text shows the relationship between the Fisher Exact Test p-values and the Jaccard index for all pathway pairs. The data shows a very strong correlation between the two (Pearson correlation index of 0.87), which confirms our hypothesis that the crosstalk can be explained by the presence of genes that are involved in more than one pathway. Very similar results have been obtained for FCS analysis (GSEA) and for the impact analysis (SPIA) (see Fig. 2 in the main text). The Pearson correlation between the p-values provided by GSEA and the Jaccard indices of all KEGG pathways was 0.62, while in the case of

SPIA the correlation was 0.83.

2 Supplementary results

Fat remodeling in mice. The results discussed here were obtained from the comparison between expression levels of genes at days 7 and 0 in the same fat remodeling experiment discussed in the main text. Genes were ordered by p-values and the top 5% were selected as differentially expressed (DE). The results of the classical ORA are shown in Fig. S3a (only the top 20 pathways are shown). The top pathways are *Parkinson's disease, Alzheimer's disease*, and *Huntington's disease*, diseases that have little to do with the tissue remodeling phenomenon. The *Cell Cycle* pathway is likely to be related to tissue remodeling (Lee et al., 2012), and *p53 Signaling* is known to be a central pathway in the response to cellular stress, including inflammation, and related to processes like cellular senescence and cell cycle (Hussain and Harris, 2006). With four false positives in the top five pathways, the results of the classical ORA are distorted by pathway crosstalk phenomena to the point of being useless.

In order to identify and eliminate the crosstalk effects we computed the *crosstalk matrix* described in the *Materials and Methods* section. Fig. S2 represents a detail of the entire matrix. The areas marked with *a* highlight the same phenomenon present in the matrix corresponding to the comparison between days 3 and 0 of the same experiment. The significance of the pathways *Parkinson's disease*, *Huntington's disease*, *Alzheimer's disease*, and *Cardiac Muscle Contraction* is entirely due to the same mitochondrial activity pathway shown in Fig. 5 in the main text. The greatly enhanced mitochondrial activity in the treated tissue was validated in vivo by in-situ hybridization (see Fig. 6 in the main text). This shows additional evidence towards the activation of this independent pathway in this condition.

We then applied the proposed Maximum Impact Estimation (fully described in Materials and Methods) to the data set. The ranking obtained with the p-values corrected for crosstalk is shown in Fig. S3b and is greatly improved. The most significant pathway is the mitochondrial pathway, showing that greatly enhanced mitochondria activity continues to be the most important difference between the treated and untreated cells even after 7 days. This in turn suggests that the tissue underwent a long-lasting remodeling phenomenon, in addition to a number of transitory phenomena such as cellular death and phagocytosis (note that the *Phagosome* pathway, significantly impacted after 3 days is not significant anymore after 7 days). The pathway ranked second is *Arrhythmogenic Right Ventricular Cardiomiopathy*. While this pathway was treated here as a false positive due to lack of literature evidence linking it specifically to tissue remodeling, the module reported by the method includes genes related to *desmosomes*, cell structures responsible for certain types of cellular adhesion (Klessner et al., 2009) which may also be relevant here. Fourth and fifth pathways in rank are, respectively, the *PPAR Signaling* pathway and the *Cell Adhesion Molecules* pathway, both closely related to the phenomenon of fat remodeling (Granneman et al., 2005).

Estrogen treatment on post-menopausal women. The second data set analyzed was produced by an experiment investigating the effect of various types of hormones on the endometrium of healthy, post-menopausal women who underwent hysterectomy (Hanifi-Moghaddam et al., 2007). Hormone therapy has been used for the treatment of conditions associated with menopause (Nelson et al., 2002). Estrogen replacement therapy has been proven useful against the insurgence of collateral effects of the post-menopausal syndrome (Campbell and Whitehead, 1977; Henderson et al., 1986; Weiss et al., 1980). However, the administration of estrogens only has been shown to increase the incidence of endometrial carcinoma (Ziel, 1982). Therefore, in addition to

estrogen, progestins are now given to menopausal women. Although the risk of endometrial cancer is reduced with the addition of progestins, the incidence of other forms of cancer seems to increase when progestin is administered with estrogen. Initiatives like the Million Women Study (http://www.millionwomenstudy.org/) and the Women Health Initiative (http://www.nhlbi.nih.gov/whi/) showed that hormone replacement therapy can increase the risk of lung and breast cancer (Chlebowski et al., 2010, 2009b). In this context, it is interesting to compare the effects of various combinations of hormones at the transcriptome level (Hanifi-Moghaddam et al., 2007).

Here, we illustrate our analysis method on the comparison of the expression levels of genes from samples treated with estrogen (E2) plus medroxyprogesterone acetate (MPA) versus normal samples. The classical over-representation analysis (ORA) finds the following pathways significant at the 5% level after FDR correction: *ECM receptor interaction, Focal Adhesion, Pathways in Cancer, Small Cell Lung Cancer, Axon Guidance, Prostate Cancer, and Jak-STAT Signaling.* These results are shown in Fig. S4a.

The E2+MPA is known to be associated with certain type of cancer including non-small-cell lung cancer (NSCLC) (Chlebowski et al., 2009a). Hence, the presence of *Pathways in Cancer* is justified, even though its identification as significant does not help understand the specific mechanism that might be active here. However, the set of significant pathways include small-cell lung cancer (SCLC) which is *not* known to be associated with this treatment and fail to include the NSCLC which *has been* linked to it (Chlebowski et al., 2009a). *Prostate Cancer* is also unlikely to be related to this specific treatment given that this treatment is administered to women, rather than men. Like in the previous case, the presence of false positives and the presence of pathways describing general cellular adhesion processes (focal adhesion and ECM-receptor interaction) does not help with the understanding of the underlying phenomenon.

After applying our analysis method to the dataset, the results are more helpful in providing insights about the specific underlying mechanisms, as shown in Fig. S4b. The first pathway in the ranked list is the *Jak-STAT signaling pathway.* Indeed, there is evidence that estrogen treatments impact such pathway through interaction with the suppressor of cytokine signaling (SOCS2) (Leung et al., 2003). The second pathway is a new pathway, based on the module common between *Focal Adhesion, ECM-receptor Interaction*, and *Pathways in Cancer* (see the left panel of Fig. S5). In this figure, within the significant quadrant, the symmetric pattern that can be observed between the three pathways above and *Pathways in Cancer* indicate the presence of a functional module that responds specifically to the hormone treatment. Interestingly, this pathway is the same pathway that has been shown to be active in a completely different phenotype, the cervical ripening experiment described in the main text, the *Integrin-Mediated ECM Signaling* described in Fig. 8. This pathway is responsible for the significance of the top four pathways in Fig. S4a.

As in the experiment studying cervical ripening, this novel pathway is composed of genes present in the interaction between the cellular transmembrane protein integrin and three important ECM components, collagen, laminin, and fibronectin, all of which appeared as differentially expressed in hormone treatment compared to the control. This is interesting because the ECM-receptor interaction carries two major functions: the first is to transduce extracellular signals into the cell for regulation of downstream pathways possibly through focal adhesion complex, and the second function is to provide structural support to resident cells; the binding between integrins and collagen, laminin, fibronectin is involved in the second process. Collagen, a major component of the ECM, forms fibers and attaches to the cell surface through binding with integrins and fibronectins. Collagen is also present in the basement membrane with laminin, forming a thin sheet of fibers that underlies the epithelium (Alberts et al., 2002). Previous studies have shown that collagen, laminin, and fibronectin participate in regulating normal development of mammalian mammary tissues (Berry et al., 2003). They also play an important role in cancer progression possibly through

ECM remodeling, which leads to alterations in cell adhesion and tumor cell motility. Consistent with this, enhanced attachment of estrogen-dependent breast cancer cells to the substrate containing ECM components (collagen I and IV, laminin, fibronectin) was observed with E2 treatment (Millon et al., 1989). More evidence was provided in recent studies using mouse mammary epithelial cells, where the expression of estrogen receptor alpha (ESR1) was greatly down-regulated by integrin-mediated interaction with collagen-IV and laminin, rather than effects of growth factors such as insulin (Novaro et al., 2003). Consistently to the previous findings, our method finds that it is the module describing the interaction between integrin and collagen, laminin, and fibronectin (rather than the interaction between ligands and their receptors) that is affected specifically by the hormone treatment, a striking pattern unlikely to be detected by classical over-representation analysis.

A similar pattern was observed between pathways *Prostate Cancer* and *Focal Adhesion*, where the removal of a common submodule caused loss of significance in both pathways. A close investigation of the *Focal Adhesion* pathway revealed that its downstream signaling cascade is regulated by two types of extracellular signals, the ECM components that interact with integrins, and the growth factors (EGF) that bind to the transmembrane GF receptor (EGFR). Although a number of DE genes belong to the *ECM-Receptor Interaction* pathway, it is the EGFR-induced signaling cascade that is involved in both *Prostate cancer* and *Focal adhesion*, which contains at least two downstream pathways that responded specifically to the E2+MPA treatment. The first one is the canonical Wnt cascade, during which the transcription factor beta-catenin gets activated by PI3K-AKT (phosphatidylinositol 3 kinase-V-Akt murine thymoma viral oncogene homolog) mediated signals, and translocate into the nucleus for downstream gene regulation (Naito et al., 2005). The other is the classical *MAPK* (Mitogen-Activated Protein Kinase) *pathway*, also known as the RAF-MAP2K-MAPK pathway, where RAF, MAP2K, and MAPK represent the three key serine/threonine-specific protein kinases present in the cascade (Zhang and Liu, 2002). What is also noticeable is that in both cases, while *Wnt Signaling* pathway and *MAPK Signaling* pathway both contain sub-pathways other than the two highlighted here, such as the *Wnt5-induced non-canonical Wnt pathway* or *JNK-p38-mediated MAPK pathway*, only the canonical Wnt cascade and the classical MAPK cascade are associated with both *Prostate Cancer* and *Focal Adhesion*, among which a number of important genes are DE under the hormone condition, such as PTEN (phosphatase and tensin homolog), a tumor suppressor that regulates PI3K-AKT signaling pathway, MAPK, one of the three key protein kinases in the MAPK pathway, and AR (androgen receptor), an oncogene that plays an important role in MAPK-regulated cell proliferation (Han et al., 2005; Peterziel et al., 1999). Indeed, estradiol has been shown to activate beta-catenin-mediated Wnt pathway through inhibition of its partner GSK3 in the rat hippocampus, which releases beta-catenin and allows its nuclear translocation (Cardona-Gomez et al., 2004). More functional evidence was provided using human colon and breast cancer cells, in which estrogen receptor (ER) and beta-catenin were found to participate in the same multi-protein complex, whose interaction gets enhanced with the presence of estrogen (Kouzmenko et al., 2004). Since both beta-catenin and ER function as transcription factors, it is possible that the role of beta-catenin in this complex is to recruit additional co-activators and chromatin remodeling factors that interact with ER for downstream transcriptional regulation (Kouzmenko et al., 2004). Estrogen has been demonstrated to induce cell proliferation through increased phosphorylation of MAPK cascade, with the mechanistic link between estrogen and MAPK signaling lying in a partner of ER, the PELP1 (proline, glutamate and leucine rich protein 1, the modulator of non-genomic activity of estrogen receptor) protein (Wong et al., 2002). PELP1 forms a complex with ER and Src family of tyrosine kinases as a scaffold protein, which is enhanced by E2, further induces activation of MAPK kinases and affects ER-mediated transcription (Wong et al., 2002). Consistent with these studies, our method detected

a module shared between *Prostate cancer* and *Focal adhesion*, the EGFR-induced canonical Wnt and classical MAPK cascade, which is responsible for significance of both pathways.

Another interesting case is shown in the right panel of Fig. S5. Here, the *Graft-Versus-Host Disease* pathway *gains* significance when the crosstalk of various other pathways is removed. This happens because all shared genes between *Graft-Versus-Host Disease* and the others are all non-DE genes in this condition. In other words, the DE genes present in *Graft-Versus-Host Disease* pathway are specific to the pathway itself. Among those, two particularly interesting ones are PRF1 (perforin 1) and GZMB (granzyme B), both of which play important functional roles in the natural killer (NK) cell-mediated cytolysis. Consistent with this, the *Graft-Versus-Host Disease* pathway is highlighted as being significantly affected by the E2+MPA treatment in the crosstalk matrix, not due to other interactions but due to *genes specific to NK cell-mediated cytotoxicity*. It is remarkable that the results of this type of analysis allowed the identification of a module, composed by genes belonging to the *Graft-Versus-Host Disease* pathway, that is impacted by the hormone treatment, and whose importance was masked by crosstalk effects with other pathways. This module is relevant in the condition studied, and treating it separately would provide a more accurate understanding of the underlying biological phenomenon. However, since the activity of this module was not identified yet in another condition, nor do we have an independent in vivo validation for this phenotype, we are not proposing this as an independent pathway at this time.

Crosstalk matrix for the cervical ripening experiment. The crosstalk matrix for the cervical ripening experiment indicates the presence of an independent functional module among the top three pathways in the ranking. The module is the same module found in the hormone treatment experiment, although in this experiment it is found from the interaction of different pathways. A detail of the crosstalk matrix is shown in Fig. S6.

Alzheimer's disease. We analyzed the data set produced by an experiment investigating the correlation between gene expression values "*with MiniMental Status Examination (MMSE) and neurofibrillary tangle (NFT)*" in subjects with Alzheimer's disease (Blalock et al., 2004). Figures S7a and S7b show the comparison between the results of the classical ORA and the results of the crosstalk analysis. At the top of the results of the ORA we find *Huntington's, Alzheimer, Parkinson's, Glutamatergic Synapse, and Arrhythmogenic right ventricular cardiomyopathy*. In this list, *Alzheimer's* is the obvious true positive, *Huntington's, Parkinson's, and Glutamatergic Synapse* are definitely related to the phenomenon, being involved in neurodegenerative diseases, while *Arrhythmogenic Right Ventricular Cardiomyopathy* is clearly a false positive. The cross-talk analysis reports, as only significant pathway, the module composed by the intersection between the *Alzheimer's, Parkinson's, and Huntington's* pathways.

The DE genes in this module consist are related to the phenomena of oxidative phosphorylation and cytochrome oxidation. There is evidence (Mecocci et al., 1994; Parker et al., 1990; Zhu et al., 2004) that these mechanisms are indeed central in Alzheimer's, and the crosstalk analysis was able to pinpoint the functional sub-pathway that is responsible for the phenotype, eliminating the false positive present in the classical analysis list.

Alzheimer's Disease - Reactome database. We analyzed the dataset produced in (Blalock et al., 2004) against the set of pathways from the Reactome database (Joshi-Tope et al., 2005). The results of the crosstalk analysis are shown in Figures S8a (for the ORA) and S8b (for the crosstalk analysis). In this case, the crosstalk analysis compacts the pathways that are at the top of the ORA result. Those pathways are all related to Alzheimer's disease (Marczynski, 1998; Parker

et al., 1994), and the crosstalk procedure of building the functional module that is involved in the phenomenon highlights the close interaction among them. The only false positive of the ORA result is *Regulation of Insulin Secretion*. This pathway describes signaling involving pancreatic beta cells and it is not related to brain cells. This pathway is not significant anymore after the correction for crosstalk. It has to be noted that there is no *Alzheimer's* specific pathway in the Reactome database. However, the crosstalk analysis was able to identify highly related pathways, providing a more concise result list with no obvious false positives.

3 Materials and methods

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 over-represented, 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, hence is devoid of crosstalk, and then to perform the ORA 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 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 $P_1 \dots P_k$. 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 $X_{i,j}$ of the matrix X is defined as follows:

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

The matrix $Y|X$ obtained by combining the vector Y with the X matrix is shown in the example in Fig. S9.

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 as described in the following.

Assuming that in a specific biological condition each gene distributes its impact differently to each pathway, we will consider the pathway to which each gene distributes the greatest fraction of its impact. We define a binary matrix Z that indicates, for each gene, the pathway that

receives the biggest fraction of that gene's impact. For each gene g_i, the corresponding row $Z_i = [Z_{i1}, Z_{i2}, \ldots, Z_{ik}]$, where $Z_{ij} \in \{0, 1\}$, will have $\sum_{j=1}^{k} Z_{ij} = 1$, i.e. there is only one column in each row that has a non-zero element. This matrix Z is the unknown underlying pathway impact matrix referred to above; our goal is to estimate it.

Let us consider one row Z_i having a one in an unknown column j and zeros elsewhere. Since we don't know j, we compute the probability of each pathway to be the one where gene g_i gives the greatest fraction of its impact. To do this, we assume a non-negative vector of multinomial probabilities $\Pi = (\pi_1, \ldots, \pi_k)$ with $\sum_{j=1}^{k} \pi_j = 1$, defined by $\pi_j = p(Z_{ij} = 1 | Y_i = 1)$. In other words, given a gene g_i that is DE, π_j is the probability that g_i gives the greatest fraction of its impact to P_j. Similarly, we also define $\Theta = (\theta_1, \ldots, \theta_k)$, where $\theta_j = p(Z_{ij} = 1 | Y_i = 0)$ for the NDE genes.

Row i of the membership matrix X is denoted by X_i; this vector tells us which pathways gene i belongs to. Within the context of the probabilistic model described above, each row X_i can be interpreted as an observation of a gene with a given expression state Y that gives the greatest fraction of its impact to one of the pathways it belongs to. Therefore, for DE genes we have $p(X_i = x_i | Y_i = 1, \Pi) = \Pi \cdot x_i'$. We further assume that the hidden matrix Z is consistent with the observed X, i.e., Z_{ij} can be 1 only when $X_{ij} = 1$; if $X_{ij} = 0$ then we must have $Z_{ij} = 0$ (a gene cannot contribute most to a pathway that it does not belong to). With this notation:

$$\begin{aligned} p(Z_i = z_i | X_i = x_i, Y_i = 1, \Pi) &= \frac{p(Z_i = z_i, X_i = x_i | Y_i = 1, \Pi)}{p(X_i = x_i | Y_i = 1, \Pi)} \\ &= \frac{I(z_i \cdot x_i' = 1) \cdot \Pi \cdot z_i'}{\Pi \cdot x_i'} \end{aligned} \tag{1}$$

where $I(.)$ is the indicator function. For example, if $x_i = (11001)$ and g_i is a DE gene, then the conditional distribution of Z_i is given by:

$$\begin{aligned} &p(Z_i = (10000) | X_i = x_i, Y_i = 1, \Pi) = \pi_1 / (\pi_1 + \pi_2 + \pi_5) \\ &p(Z_i = (01000) | X_i = x_i, Y_i = 1, \Pi) = \pi_2 / (\pi_1 + \pi_2 + \pi_5) \\ &p(Z_i = (00100) | X_i = x_i, Y_i = 1, \Pi) = 0 \\ &p(Z_i = (00010) | X_i = x_i, Y_i = 1, \Pi) = 0 \\ &p(Z_i = (00001) | X_i = x_i, Y_i = 1, \Pi) = \pi_5 / (\pi_1 + \pi_2 + \pi_5) \end{aligned} \tag{2}$$

This yields a vector of conditional probabilities $c_i = (c_{i1}, c_{i2}, ..., c_{ik})$ for each row Z_i of DE genes, where $c_{ij} = p(Z_{ij} = z_{ij} |\ X_i = x_i)$ as defined above. Once those probabilities are estimated, we can produce a most likely matrix Z by assigning each gene to the pathway with the highest probability of receiving the biggest fraction of the impact of the gene. Specifically, $z_{ij} = 1$ when $\max_s\{c_{is}\} = c_{ij}$; $z_{ij} = 0$ otherwise.

If there were no crosstalk, each gene would contribute to a single pathway, the matrix X and the matrix Z would be equal, and they would have only one element equal to 1 in each row. In this case, π_j could be estimated as the number of DE genes belonging to the pathway divided by the total number of DE genes. The probabilities π and θ could be estimated as follows:

$$\widehat{\pi}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{3}$$

$$\widehat{\theta}_j = \frac{\sum_{i=n+1}^{n+m} x_{ij}}{m} \tag{4}$$

In the presence of crosstalk, however, it is not possible to compute Π and Θ directly from X. A likelihood-based estimation can be used instead.

The log-likelihood of observing the membership matrix X given the gene expression vector Y is then:

$$logL = \sum_{i=1}^{n+m} log(p(X_i|Y_i; \pi_1, \pi_2, \pi_3 \dots \pi_k, \theta_1, \theta_2, \theta_3 \dots \theta_k)) \tag{5}$$

Equation 5 is written under the assumption of conditional independence of rows of X; i.e., under the reasonable assumption that the pathway to which a gene i gives most of its impact does not depend on the pathway to which another gene j impacts the most. In other words, the split of the fractions of the impact of a gene does not depend the splits of the impact of other genes.

This assumption, together with the observation that the DE genes do not depend on θ's and that the NDE genes do not depend on π's, allows us to compute the likelihood by separating the matrix in two sub-matrices: $X|Y = 1$, representing the sub-matrix of the DE genes, and $X|Y = 0$, representing the sub-matrix of the NDE genes:

$$\begin{aligned} logL &= \sum_{i=1}^{n} log(p(X_i|Y_i = 1, \Pi)) + \sum_{i=n+1}^{m+n} log(p(X_i|Y_i = 0, \Theta)) \\ &= \sum_{i=1}^{n} log(\Pi \cdot X_i') + \sum_{i=n+1}^{m+n} log(\Theta \cdot X_i') \end{aligned} \tag{6}$$

In this formula, the (row) vector Π represents the probability of the $i-th$ DE gene to give the greatest fraction of its impact to a specific pathway, X_i is the $i-th$ row of the membership matrix X, and X_i' represents its transpose. The dot-product $\Pi \cdot X_I'$ produces a scalar representing the probability $P(X_i = x_i|Y = 1, \Pi)$, i.e. the probability of observing the i-th row of the matrix X_i given the fact that gene i is DE. The same notation has been used for the dot-product $\Theta \cdot X_i'$.

In the following, we will only work with the first term to illustrate how to estimate Π. Θ can be estimated from $X|Y = 0$ in a similar fashion.

There is no closed form solution for the maximization of Eq. 6. However, we can use the Z matrix as a hidden variable for the estimation of the parameters Π. The log joint conditional likelihood for the DE part of the matrix can be written as:

$$\begin{aligned} logJL^{DE} &= log(p(X, Z|Y = 1, \Pi)) \\ &= \sum_{i=1}^{n} log(p(X_i, Z_i|Y_i = 1, \Pi)) \\ &= \sum_{i=1}^{n} log(I(Z_i^{DE} \cdot (X_i^{DE})' = 1) \cdot Z_i^{DE} \cdot \Pi) \\ &= \sum_{i=1}^{n} \left(log(I(Z_i^{DE} \cdot (X_i^{DE})' = 1)) \cdot \sum_{j=1}^{k} z_{i,j}^{DE} \cdot log(\pi_j) \right) \\ &= \sum_{i=1}^{n} log(\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE}) + \sum_{i=1}^{n} \sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE} \end{aligned} \tag{7}$$

We use an expectation maximization (EM) approach to maximize the log likelihood in Equation 5 by maximizing the joint log likelihood defined in Equation 7. The EM is an iterative algorithm that starts with an initial guess for Π, denoted with Π^0; each iteration is a mapping between Π^t and Π^{t+1} .The superscript indicates the index of the iteration. We choose to initialize each element of the vector as follows:

$$\pi_j^0 = \frac{\sum_{i=1}^{n} x_{i,j}}{\sum_{i=1}^{n}\sum_{h=1}^{k} x_{i,h}},\ j \in \{1 \dots k\} \tag{8}$$

This initializes each value π_j with the ratio between the number of DE genes in pathway j and the sum over the matrix X. This initialization is consistent with the model described in Equation 3.

Each iteration of the EM algorithm is composed by two steps: the expectation step and the maximization step; during the expectation step we compute the expectation of the log joint conditional likelihood in Equation 7 with respect to the posterior $p(Z_{i,j}^{DE}|X_i^{DE}, \Pi^{old})$:

$$\begin{aligned}
&E\left(\sum_{i=1}^{n} log(\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE}) + \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}\right) \\
&= E\left(\sum_{i=1}^{n} log(\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE})\right) + E\left(\sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}\right) \\
&= E\left(\sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}\right)
\end{aligned} \tag{9}$$

The term $E\left(\sum_{i=1}^{n} log(\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE})\right)$ is equal to 0 because the term $\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE}$ is equal to 1 for the consistency of Z with X.

The derivation of the non zero term of the expectation is as follows:

$$\begin{aligned}
E\left(\sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}\right) &= \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot E\left(z_{i,j}^{DE}|X_{i,j}^{DE}, \Pi^{old}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot p(z_{i,j}^{DE} = 1|X_i^{DE}, \Pi^{old}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot \frac{p(z_{i,j}^{DE}, X_i^{DE}|\Pi^{old})}{\sum_{r=1}^{k} p(z_{i,j}^{DE}, X_i^{DE}|\Pi^{old})} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot \frac{x_{i,j}^{DE} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,j}^{DE} \cdot \pi_r^{old}}
\end{aligned} \tag{10}$$

The maximization of the expectation with respect to Π, subject to the constraint that $\sum_{j=1}^{k} \pi_j = 1$, is obtained with the Lagrange multiplier method as follows:

$$\frac{d[\sum_{j=1}^{k} log(\pi_j) \sum_{i=1}^{n} \frac{x_{i,h} \cdot \pi_h^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} + \lambda((\sum_{j=1}^{k} \pi_j) - 1)]}{d\pi_h} = 0, \forall h \in \{1 \dots k\}$$

$$\frac{\sum_{i=1}^{n} \frac{x_{i,h} \cdot \pi_h^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_h} + \lambda = 0, \forall h \in \{1 \dots k\} \tag{11}$$

We can write a systems of equations over all the possible values of h in order to compute λ.

$$\begin{cases} \frac{\sum_{i=1}^{n} \frac{x_{i,1} \cdot \pi_1^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r}}{\pi_1}^{old} + \lambda & = 0 \\ \vdots & \\ \frac{\sum_{i=1}^{n} \frac{x_{i,k} \cdot \pi_k^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_k} + \lambda & = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} \frac{x_{i,1} \cdot \pi_1^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} & = -\lambda \cdot \pi_1 \\ \vdots & \\ \sum_{i=1}^{n} \frac{x_{i,k} \cdot \pi_k^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} & = -\lambda \cdot \pi_k \end{cases}$$

Summing left and right sides we obtain:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} = -\lambda \sum_{j=1}^{k} \pi_j \tag{12}$$

Since $\sum_{j=1}^{k} \pi_j = 1$, we can write:

$$\lambda = -\sum_{j=1}^{k} \sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} \tag{13}$$

We substitute λ in 11 and use an iterative process in which a new π value is calculated at each step:

$$
\begin{aligned}
\frac{\sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_h^{new}} + \lambda = 0, \forall h \in \{1 \dots k\} \\
\frac{\sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_h^{new}} - \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} = 0, \forall h \in \{1 \dots k\} \\
\frac{\sum_{i=1}^{n} \frac{x_{i,h} \cdot \pi_h^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_h^{new}} = \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}, \forall h \in \{1 \dots k\} \\
\pi_h^{new} = \frac{\sum_{i=1}^{n} \frac{x_{i,h} \cdot \pi_h^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\sum_{j=1}^{k} \sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}, \forall h \in \{1 \dots k\}
\end{aligned}
\tag{14}
$$

Since the sum over each row is 1, if we invert the order of the summations at the denominator in the last row of Equation 14, the value of the denominator becomes n. This, in other words, means that each value π_h^{new} is the sum of column h over the number of DE genes.

The algorithm stops when the distance between two consecutive vectors $||\Pi^{(t)} - \Pi^{(t-1)}||$ is less than the quantity $\frac{||\Pi^{(1)} - \Pi^{(0)}||}{100}$, i.e. the distance between the first two vectors divided by 100. At the end of the steps of the EM algorithm we obtain the matrix C from which we can obtain the most probable Z given the condition under study: for each row, we assign the value 1 to the cell with the highest probability, and 0 to all the others. This is equivalent to saying that each gene gives its full impact to the pathway with the highest π value.

Module Detection. The module detection procedure uses an iterative approach for the identification of independent functional modules. At each iteration, only one module is tested, for example the module in common between pathways P_i and P_j. The set of pathways over which we perform the correction for multiple comparison is the set of original pathways, without the two pathways P_i and P_j, and with the inclusion of the pathways $P_{i \setminus j}$, $P_{j \setminus i}$, and the module $P_{i \bigcap j}$. If the original list of pathways contained k pathways, the correction of the significance of the module is computed on a list that contains $k + 1$ pathways.

It has to be noted that the goal of the module detection process is not to compute the exact significance of each module, but to estimate the change of the significance of a pair of pathways when the intersection among them is removed. If this change is big enough, and the intersection's significance is comparable to the one of the original pathways, we assume that the module is the responsible for the significance of the two parent pathways, and we modify the list of pathways accordingly. When the list of pathways is modified with the addition of the newly discovered modules, the correction for multiple comparisons is performed on the new augmented list, estimating the significance of pathways and modules appropriately. If there are n new modules added to the original list of k pathways, there will be $k+n$ tests and we do correct for $k+n$ multiple comparisons.

Choice for the threshold for the module detection procedure. The value 0.25 for the module selection procedure was selected by calculating all modules for all data sets with different thresholds in the $[0, 0.4]$ range (with a difference of 0.025 between thresholds). The results are

shown in Figure S10. As it can be seen in the figure, the number of modules found in all data sets shows a plateau in the $[0.1, 0.375]$ range.

The false negative rate of the crosstalk correction method. We ran the crosstalk correction method on the simulations described in Section 1 and we computed the number of times that the crosstalk correction considers the bait pathway as significant, giving an estimation of the detection power of the method. In simulations that were designed to yield 100% power for the Fisher Exact Test, the crosstalk correction reported the bait as significant 91.5% of the time (yielding a 8.5% false negative rate). This loss of power is not excessive, and is the trade-off for eliminating crosstalk effects. Furthermore, the 91.5% power is well above the practical threshold of 80% commonly used in randomized clinical trials (Cohen, 1988; Ellis, 2010; Hulley et al., 1998).

Pathway size bias analysis. To evaluate the possible effect of the pathway size on the results of the crosstalk correction, we produced histograms of the p-values for different ranges of pathway sizes, as well as scatter plots representing the crosstalk corrected p-values versus pathway size.

Figure S11 shows the histograms. The pathways have been divided into quartiles by the size of the pathways. The quartiles are shown from the top (first quartile) to the bottom (last quartile). The cyan bars (to the left in each interval) represent the p-values yielded by the crosstalk correction, the red bars represent the p-values yielded by the Fisher Exact Test. The data shows that the crosstalk p-values are distributed very similarly to the Fisher Exact Test p-values. Furthermore, there are no major differences between the groups of pathways with different sizes.

Figure S12 shows the scatter plots. The top two panels in this figure show the crosstalk corrected p-values (panel **(a)**) and the p-values of the Fisher Exact Test (panel **(b)**). The empty spaces between the curves outlined by the points are due to the discrete nature of the hypergeometric distribution. Since the number of DE genes only takes natural values, the p-values of any pathway will have discrete values for any given pathway size (going vertically for any value of x). In order to illustrate this, we colored the points corresponding to a number of DE genes in the interval $[1, 9]$. Since there are many points close to 0 and many points close to 1, a potential size bias may not be apparent. Hence, we calculated the Pearson correlation between the p-values and the sizes of the pathways. The correlation between the crosstalk corrected p-values and the pathway sizes is -0.052; the correlation between the the Fisher Exact Test p-values and the pathway sizes is -0.079.

Figure legends

Figure S1: The distributions of the p-values obtained from the three analysis methods under the null hypothesis: Fisher's Exact Test (left), SPIA (middle), and GSEA (right). All three exhibit a significant departure from the expected uniform distribution (Kolmogorov-Smirnov p-values of the order of 10^{-16} in all cases). Notably, all methods yield a much higher than expected number of pathways with p-values lower than 0.1, i.e. false positives.

Figure S2: Detail of the crosstalk matrix for the comparison between days 7 and 0 in the fat remodeling experiment. The areas marked with *a* correspond to the *Mitochondrial activity* pathway shown in Fig. 5, the same pathway that was found to be activated in the dataset associated with the comparison of expression levels at days 3 and 0.

(a) The top 20 pathways resulting from classical ORA before correction for crosstalk. Four out of the top five pathways are not related to the fat remodeling phenomenon (false positives).

(b) The top 20 pathways after the correction for crosstalk effects. The mitochondrial activity pathway (validated in vivo) is reported as the most significant pathway even after 7 days, suggesting permanent tissue remodeling. The *Phagosome* pathway, significantly impacted after 3 days (see Fig. 3b) is not significant anymore after 7 days, consistent with the transitory nature of cellular death and phagocytosis. The four false positives in the left table have been removed. The ARVC is reported as a false positive but the DE genes located on this pathway are involved in cell adhesion which may be relevant here.

Figure S3: The results of the ORA for the fat remodeling experiment for the comparison between days 7 and 0, before (left) and after (right) the correction for crosstalk effects. All p-values are FDR-corrected. The lines show the significance thresholds: blue - 0.01, yellow - 0.05.

(a) The top 20 pathways reported by the classical ORA before correction for crosstalk. The NSCLC, known to be linked to this treatment (Chlebowski et al., 2009a) is not identified by the classical method, while the SCLC, which showed no increase in incidence in the treatment group (Chlebowski et al., 2009a), appears as significant. The significance of *Pathways in Cancer* is consistent with the putative link between hormone treatments and higher incidence of some types of cancer but offers no explanation or insight into the underlying mechanisms.

(b) The top 20 pathways reported by ORA after the correction for crosstalk effects. The correction method removed *Pathways in Cancer*, *SCLC*, and *Prostate Cancer* from the list of significant pathways, increasing the significance of pathways offering more insights such as *Jak-STAT signaling pathway* and the new *Integrin mediated ECM signaling* module. A star before the name of the pathway means that a module overlapping with other pathways has been removed from the pathway.

Figure S4: Results of ORA for the estrogen treatment experiment, before (left) and after (right) the correction for crosstalk effects. All p-values are FDR-corrected. The lines show the significance thresholds: blue - 0.01, yellow - 0.05.

Figure S5: Detail of the crosstalk matrix of the estrogen treatment. Left panel: the circle highlights an example of a common module that is responsible for the significance of an entire group of pathways. The common module between the pathways *ECM-Receptor Interaction, Focal Adhesion, Pathways In Cancer*, and *Small Cell Lung Cancer* describes the interaction between *integrin* and *collagen, laminin,* and *fibronectin*. Henceforth, we will refer to this module as the *Integrin-mediated ECM signaling* pathway (see Fig. 8). Right panel: row corresponding to the pathway *Graft-Versus-Host disease*. The pathway becomes significant after the removal of specific pathways, highlighted by the yellow circles. The set of pathways includes *Phagosome, Cell adhesion molecules (CAMs)*, *Leishmaniasis*, *Intestinal immune network for IgA production, Systemic Lupus Erithematosus*, and *Asthma*. This indicates a situation in which the genes specific to *Graft-Versus-Host disease* are related to the phenomenon in analysis, but their significance is *masked* by the presence of crosstalk with other pathways.

Figure S6: Details of the crosstalk matrix of the cervical ripening experiment. The circle highlights the evidence for an independent module involving pathways *Focal Adhesion, ECM-Receptor Interaction*, and *Amoebiasis*. This module is exactly the *Integrin-mediated ECM signaling* previously identified in the hormone treatment experiment (Fig 8) from a different set of crosstalk interactions. The bright green loss of significance of *Small-Cell Lung Cancer* in columns 1-3, shows that this pathway was a false positive in the ORA since its significance was due only to the crosstalk from the first 3 pathways.

(a) Results of the ORA analysis of the GSE1297 data set using KEGG as a reference database. While related to neurodegenerative diseases, the pathways Huntington's and Parkinson's are not true positives. The pathway *Arrhythmogenic right ventricular cardiomyopathy* is not related to the phenomenon.

(b) Results of the crosstalk analysis of the GSE1297 data set using KEGG as a reference database. The crosstalk analysis is able to extract a functional module from the three neurodegenerative disease pathways that rank at the top of the ORA list. Genes found in this module are related to the phenomena of oxidative phosphorylation and cytochrome oxidase, highly related to Parkinson's disease. The pathway *Arrhythmogenic right ventricular cardiomyopathy* is not significant anymore.

Figure S7: The results of the ORA analysis in the GSE1297 experiment before (left) and after (right) correction for crosstalk effects. All p-values are FDR-corrected. The blue line shows the 0.05 significance threshold.

(a) Results of the ORA analysis of the GSE1297 data set using Reactome as a reference database. The top pathways are related to Alzheimer's Disease. The pathway *Regulation of Insulin Secretion* describe the signaling events involving pancreatic beta cells, and it is not related to brain cells.

(b) Results of the crosstalk analysis of the GSE1297 data set using Reactome as a reference database. The crosstalk analysis groups the pathways related to Alzheimer's Disease. The pathway *Regulation of Insulin Secretion* is not significant anymore.

Figure S8: The results of the ORA analysis in the GSE1297 experiment using Reactome as reference database, before (left) and after (right) correction for crosstalk effects. All p-values are FDR-corrected. The blue line shows the significance thresholds of 0.05.

Figure S9: Example of a DE/membership matrix; the 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_j represents the membership indicator for pathway j. Row g_i describes gene i in terms of its differential expression and its membership to the various pathways.

Figure S10: Number of modules obtained when changing the threshold distance under which two modules are considered similar enough to be joined. All data sets showed a plateau in the $[0.1, 0.375]$ range indicating that the number of modules found does not depend on the choice of the threshold for a wide range of threshold values.

Figure S11: Histograms showing a comparison of the frequencies of p-values between the crosstalk and Fisher's Exact test. The pathways have been divided into quartiles by the size of the pathways. The quartiles are shown from the top (first quartile) to the bottom (last quartile). The cyan bars (to the left in each interval) represent the p-values yielded by the crosstalk correction, the red bars represent the p-values yielded by the Fisher Exact Test. The histograms are normalized such that the area under each of them is 1. The data shows that the cross-talk p-values are distributed very similarly to the Fisher Exact Test p-values. Furthermore, there is no evidence of any size bias (no major difference between the groups of pathways with different sizes).

Figure S12: The scatter plot showing the p-values (vertical axis) versus pathway size (horizontal axis) for crosstalk (panel a) and Fisher Exact test (panel b). The empty space is due to the fact that there are no pathways with sizes between 365 and 984. The p-values yielded by a number of DE genes from 1 to 9 have been colored.

Figures

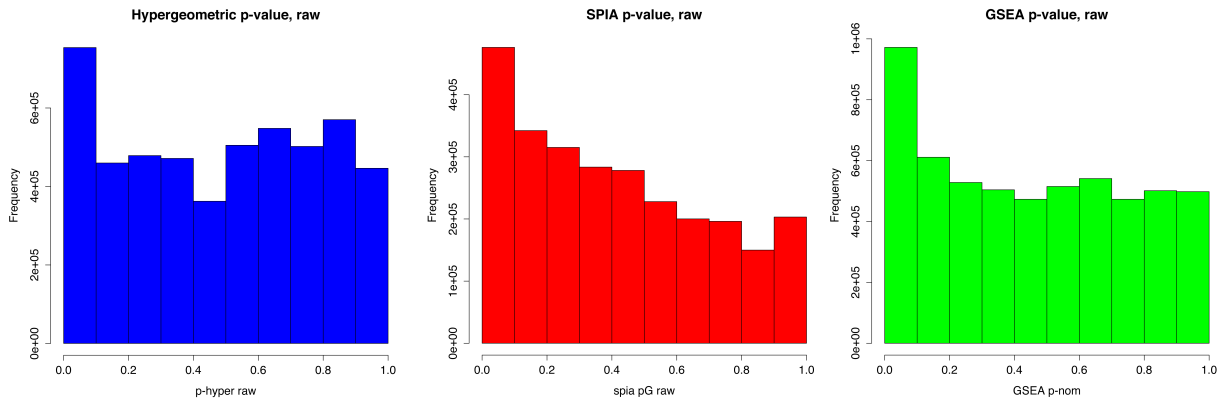

Figure S1

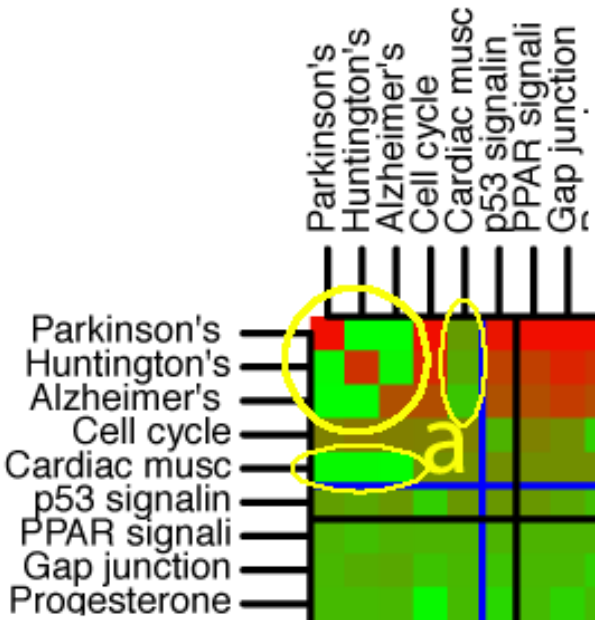

Figure S2

rank	pathway	p(FDR)
1	Parkinson's disease	$7.2e{-}06$
2	Huntington's disease	$4.2e{-}05$
3	Alzheimer's disease	0.0002
4	Cell cycle	0.0044
5	Cardiac muscle contraction	0.0087
6	p53 signaling pathway	0.0134
7	PPAR signaling pathway	0.0773
8	Gap junction	0.0920
9	Progest. mediated oocyte matur.	0.0995
10	Oocyte meiosis	0.1327
11	Salivary secretion	0.1442
12	Cell adhesion molecules (CAMs)	0.2390
13	SNARE inter. in vesicular transp.	0.2969
14	Prostate cancer	0.3837
15	Vasopressin-reg. water reabs.	0.5111
16	Arrhythm. right ventr. card.	0.5111
17	Hedgehog signaling pathway	0.5174
18	Prion diseases	0.5420
19	Melanogenesis	0.5432
20	Pathways in cancer	0.5432

(a)

rank	pathway	p(FDR)
1	Mitochondrial Activity	$2.3e{-}08$
2	Arr. right ventr. cardiom. (ARVC)	0.001
3	Cell cycle	0.001
4	PPAR signaling pathway	0.015
5	Cell adhesion molecules (CAMs)	0.019
6	Melanogenesis	0.019
7	Vascular smooth muscle contr.	0.080
8	p53 signaling pathway	0.125
9	Pathways in cancer	0.562
10	SNARE inter. in vesicular transp.	0.562
11	Chagas disease	0.575
12	Long-term potentiation	0.575
13	Phagosome	0.588
14	Vasopressin-reg. water reabs.	0.765
15	Hedgehog signaling pathway	0.765
16	Dorso-ventral axis formation	0.765
17	Intest. imm, netw. for IgA prod.	0.784
18	Wnt signaling pathway	0.984
19	ECM-receptor interaction	0.984
20	Phototransduction	0.984

(b)

Figure S3

rank	pathway	p(FDR)
1	ECM-rec. interaction	0.0343
2	Focal adhesion	0.0401
3	Pathways in cancer	0.0401
4	Small cell lung cancer	0.0401
5	Axon guidance	0.0401
6	Prostate cancer	0.0401
7	Jak-STAT signaling pathway	0.0401
8	Progest.-med. oocyte mat.	0.0951
9	Adipocytokine signaling	0.0951
10	Melanoma	0.1208
11	Graft-versus-host disease	0.1291
12	Reg. of actin cytoskeleton	0.2020
13	Aldosterone-reg. Na reabs.	0.2020
14	Oocyte meiosis	0.2168
15	Long-term depression	0.2174
16	mTOR signaling pathway	0.3048
17	Nat. killer cell med. cytotox.	0.3185
18	Vibrio cholerae infection	0.3225
19	SNARE inter. in vesicular trans.	0.3699
20	Salivary secretion	0.3699

(a)

rank	pathway	p(FDR)
1	Jak-STAT signaling pathway	$5e{-}09$
2	Integrin Mediated ECM Sign.	0.0001
3	Axon guidance	0.0036
4	Vascular sm. muscle contr.	0.0070
5	Aldosterone-reg. Na reabs.	0.0190
6	Adipocytokine signaling	0.0326
7	Nat. killer cell med. cytotox.	0.0344
8	Regulation of actin cytosk.	0.1403
9	Compl. and coag. cascades	0.3413
10	Adherens junction	0.3413
11	SNARE interac. in ves. trans.	0.4842
12	Circadian rhythm - mammal	0.5074
13	Lysosome	0.6552
14	Protein proc. in endopl. ret.	0.7182
15	Vibrio cholerae infection	0.7182
16	* * * Focal adhesion	0.9844
17	Type I diabetes mellitus	1
18	Phagosome	1
19	Huntington's disease	1
20	Cell cycle	1

(b)

Figure S4

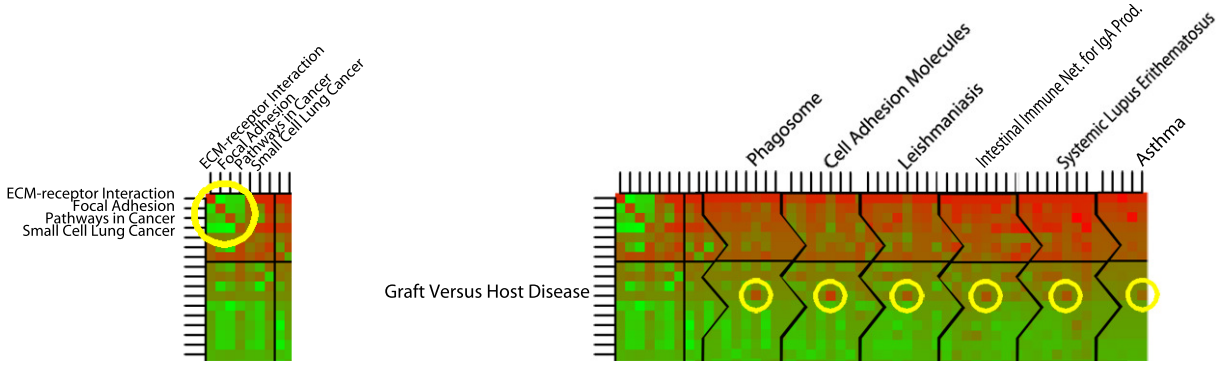

Figure S5

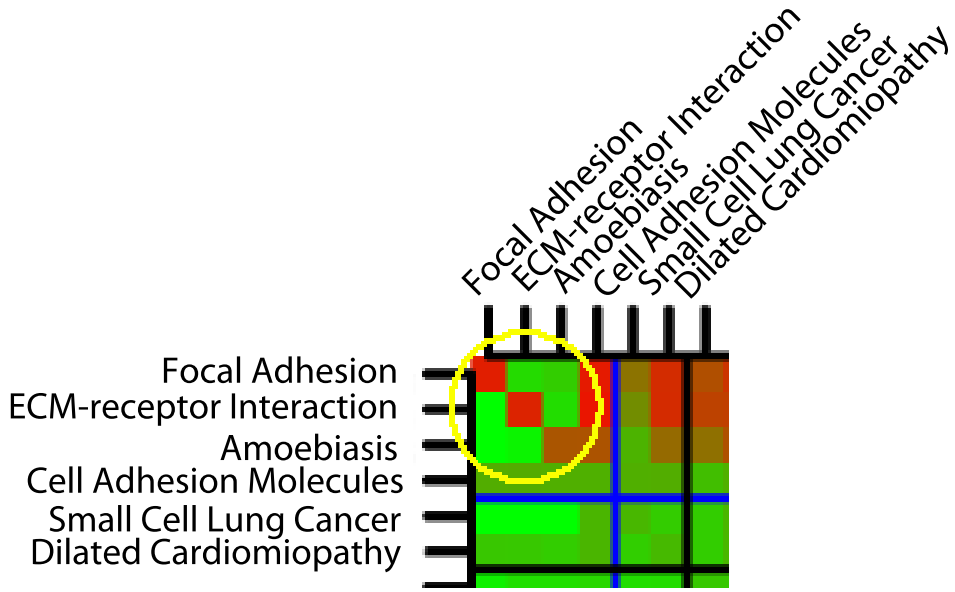

Figure S6

Rank	Title	p-value(fdr)
1	Huntington's disease	$3.49 \cdot 10^{-06}$
2	Alzheimer's disease	$3.49 \cdot 10^{-06}$
3	Parkinson's disease	$3.49 \cdot 10^{-06}$
4	Glutamatergic synapse	0.00342933
5	Arrhythm. right ventr. cardiom.	0.0110
6	Circadian rhythm - mammal	0.0995
7	Dopaminergic synapse	0.1322
8	Long-term depression	0.1625
9	Calcium signaling pathway	0.1922
10	Retrograde endocann. signaling	0.1922

(a)

Rank	Title	p-value (fdr)
1	Alzheime+Parkinso+Huntingt	$2.44 \cdot 10^{-07}$
2	Arrhythm. right ventr. cardiom.	0.1826
3	Glutamatergic synapse	0.3789
4	GABAergic synapse	0.6135
5	ECM-receptor interaction	0.6135
6	Circadian rhythm - mammal	0.6135
7	Gap junction	0.8626
8	Phosphat. signaling system	1
9	Axon guidance	1
10	Serotonergic synapse	1

(b)

Figure S7

Rank	Title	p-value (fdr)
1	Tca Cycle/Respiratory Electron Transport	$5.22 \cdot 10^{-09}$
2	Respiratory Electron/Atp Synthesis	$1.92 \cdot 10^{-07}$
3	Respiratory Electron Transport	$2.94 \cdot 10^{-05}$
4	Gaba Synthesis Release Reuptake and Degradation	$2.94 \cdot 10^{-05}$
5	Neurotr. Release Cycle	$2.94 \cdot 10^{-05}$
6	Neuronal System	$2.94 \cdot 10^{-05}$
7	Glutamate Neurotr. Release Cycle	0.0006
8	Transmission Across Chemical Synapses	0.0006
9	Formation of Atp by Chemiosm. Coup.	0.0143
10	Regulation of Insulin Secretion	0.0160
11	Norepinephrine Neurotr. Rel. Cycle	0.0160
12	Protein Folding	0.0183
13	Integration of Energy Metabolism	0.0184
14	Prefoldin Mediated Transfer of Substrate to Cct Tric	0.033
15	Darpp 32 Events	0.0374

(a)

Rank	Title	p-value (fdr)
1	Respiratory electron/atp synthesis + Respiratory electron tr. + Tca Cycle	$3.85 \cdot 10^{-07}$
2	Gaba Synth. + Glutamate neuron. + Neurona system + Neurotrans. + Transm. Chemical Synapses	0.0001
3	* Tca Cycle and Respiratory Electron Transport	0.0004
4	Prefoldi + Protein	0.1601
5	Glucose Metabolism	0.7029
6	Hemostasis	1
7	Metabolism of Nucleotides	1
8	Nuclear Signaling by Erbb4	1
9	Biological Oxidations	1
10	* Neuronal System	1
11	Axon Guidance	1
12	Smooth Muscle Contraction	1
13	G Alpha Z Signalling Events	1
14	Mitotic G2 G2 M Phases	1
15	Base Free Sugar Phosphate Removal	1

(b)

Figure S8

	Y	P_1	P_2	P_3	...	P_k
g_1	1	0	1	1	$\cdots$	0
g_2	1	0	1	0	$\cdots$	0
g_3	1	1	0	0	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
g_{n-1}	1	0	0	1	$\cdots$	0
g_n	1	0	1	0	$\cdots$	0
g_{n+1}	0	0	0	1	$\cdots$	0
g_{n+2}	0	1	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
g_{n+m-1}	0	1	0	0	$\cdots$	0
g_{n+m}	0	0	0	0	$\cdots$	0

Figure S9

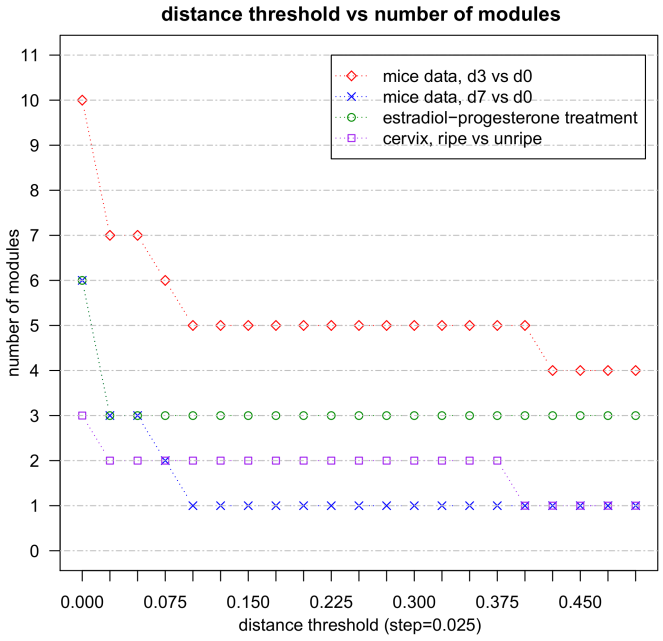
distance threshold vs number of modules
mice data, d3 vs d0
mice data, d7 vs d0
estradiol−progesterone treatment
cervix, ripe vs unripe
number of modules
0
1
2
3
4
5
6
7
8
9
10
11
0.000
0.075
0.150
0.225
0.300
0.375
0.450
distance threshold (step=0.025)

Figure S10

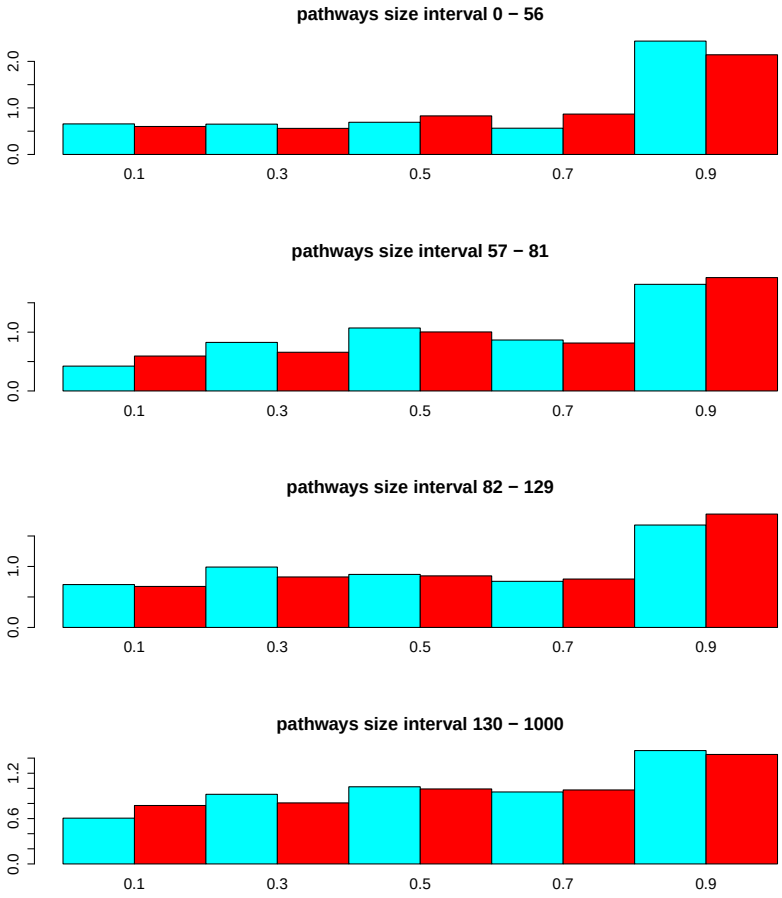
pathways size interval 0 − 56
0.0
1.0
2.0
0.1
0.3
0.5
0.7
0.9
pathways size interval 57 − 81
0.0
1.0
0.1
0.3
0.5
0.7
0.9
pathways size interval 82 − 129
0.0
1.0
0.1
0.3
0.5
0.7
0.9
pathways size interval 130 − 1000
0.0
0.6
1.2
0.1
0.3
0.5
0.7
0.9

Figure S11

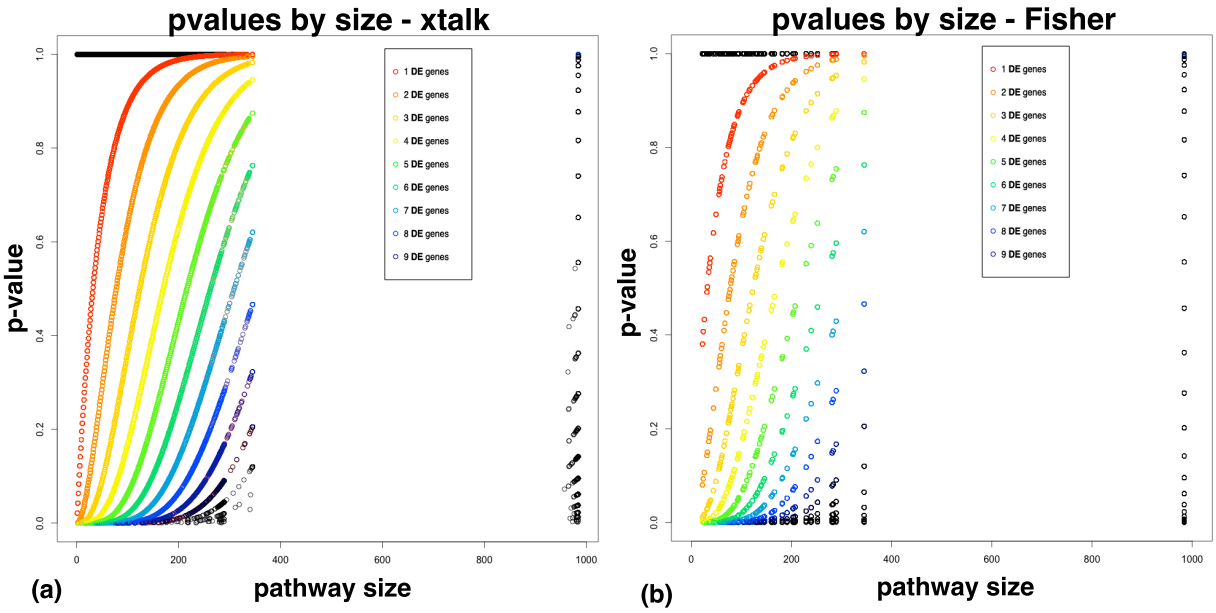
pvalues by size - xtalk
p-value
pathway size
(a)
1 DE genes
2 DE genes
3 DE genes
4 DE genes
5 DE genes
6 DE genes
7 DE genes
8 DE genes
9 DE genes
pvalues by size - Fisher
p-value
pathway size
(b)
1 DE genes
2 DE genes
3 DE genes
4 DE genes
5 DE genes
6 DE genes
7 DE genes
8 DE genes
9 DE genes

Figure S12

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., 2002. Molecular biology of the cell 4th edition. *Garland Science*, .

Berry, S. D., Howard, R. D., and Akers, R. M., 2003. Mammary localization and abundance of laminin, fibronectin, and collagen iv proteins in prepubertal heifers. *J Dairy Sci*, **86**(9):2864–74.

Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery, W. R., and Landfield, P. W., 2004. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(7):2173–2178.

Campbell, S. and Whitehead, M., 1977. Estrogens for menopausal flushing. *British Medical Journal*, **1**(6053):104–105.

Cardona-Gomez, P., Perez, M., Avila, J., Garcia-Segura, L. M., and Wandosell, F., 2004. Estradiol inhibits GSK3 and regulates interaction of estrogen receptors, GSK3, and beta-catenin in the hippocampus. *Mol Cell Neurosci*, **25**(3):363–73.

Chlebowski, R. T., Anderson, G. L., Manson, J. E., Schwartz, A. G., Wakelee, H., Gass, M., Rodabough, R. J., Johnson, K. C., Wactawski-Wende, J., Kotchen, J. M., *et al.*, 2010. Lung cancer among postmenopausal women treated with estrogen alone in the Women's Health Initiative randomized trial. *Journal of the National Cancer Institute*, **102**(18):1413–1421.

Chlebowski, R. T., Schwartz, A., Wakelee, H., Anderson, G. L., Stefanick, M. L., Manson, J. E., Chien, J. W., Chen, C., Wactawski-Wende, J., and Gass, M., *et al.*, 2009a. Non-small cell lung cancer and estrogen plus progestin use in postmenopausal women in the Women's Health Initiative randomized clinical trial – Chlebowski et al. 27 (18): CRA1500 – ASCO Meeting Abstracts. *Journal of Clinical Oncology*, **27**(185).

Chlebowski, R. T., Schwartz, A. G., Wakelee, H., Anderson, G. L., Stefanick, M. L., Manson, J. E., Rodabough, R. J., Chien, J. W., Wactawski-Wende, J., Gass, M., *et al.*, 2009b. Oestrogen plus progestin and lung cancer in postmenopausal women (Women's Health Initiative trial): a post-hoc analysis of a randomised controlled trial. *Lancet*, **374**(9697):1243–1251.

Cohen, J., 1988. *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Ellis, P. D., 2010. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Granneman, J. G., Li, P., Zhu, Z., and Lu, Y., 2005. Metabolic and cellular plasticity in white adipose tissue I: effects of beta3-adrenergic receptor activation. *American Journal Of Physiology-Endocrinology And Metabolism*, **289**(4):E608–616.

Han, G., Buchanan, G., Ittmann, M., Harris, J. M., Yu, X., Demayo, F. J., Tilley, W., and Greenberg, N. M., 2005. Mutation of the androgen receptor causes oncogenic transformation of the prostate. *Proc Natl Acad Sci U S A*, **102**(4):1151–6.

Hanifi-Moghaddam, P., Boers-Sijmons, B., Klaassens, A. H. A., van Wijk, F. H., den Bakker, M. A., Ott, M. C., Shipley, G. L., Verheul, H. A. M., Kloosterboer, H. J., Burger, C. W., *et al.*, 2007. Molecular analysis of human endometrium: short-term tibolone signaling differs significantly

from estrogen and estrogen plus progestagen signaling. *Journal of Molecular Medicine-jmm*, **85**(5):471–480.

Henderson, B. E., Ross, R. K., Paganinihill, A., and Mack, T. M., 1986. Estrogen use and cardiovascular-disease. *American Journal of Obstetrics and Gynecology*, **154**(6):1181–1186.

Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B., Vittinghoff, E., et al., 1998. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA: the journal of the American Medical Association*, **280**(7):605–613.

Hussain, S. P. and Harris, C. C., 2006. p53 biological network: at the crossroads of the cellular-stress response pathway and molecular carcinogenesis. *Journal of Nihon Medical School*, **73**(2):54–64.

Joshi-Tope, G., Gillespie, M., Vasrik, I., D'Eustachio, P., Schmidt, E., de Bone, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., *et al.*, 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, **33**(Database issue):D428–432.

Klessner, J. L., Desai, B. V., Amargo, E. V., Getsios, S., and Green, K. J., 2009. EGFR and ADAMs cooperate to regulate shedding and endocytic trafficking of the desmosomal cadherin desmoglein 2. *Molecular biology of the cell*, **20**(1):328–337.

Kouzmenko, A. P., Takeyama, K., Ito, S., Furutani, T., Sawatsubashi, S., Maki, A., Suzuki, E., Kawasaki, Y., Akiyama, T., Tabata, T., *et al.*, 2004. Wnt/beta-catenin and estrogen signaling converge in vivo. *J Biol Chem*, **279**(39):40255–8.

Lee, Y.-H., Petkova, A. P., Mottillo, E. P., and Granneman, J. G., 2012. In vivo identification of bipotential adipocyte progenitors recruited by Beta3-adrenoceptor activation and high-fat feeding. *Cell Metabolism*, **15**(4):480–491.

Leung, K. C., Doyle, N., Ballesteros, M., Sjogren, K., Watts, C. K. W., Low, T. H., Leong, G. M., Ross, R. J. M., and Ho, K. K. Y., 2003. Estrogen inhibits GH signaling by suppressing GH-induced JAK2 phosphorylation, an effect mediated by SOCS-2. *PNAS*, **100**(3):1016–1021.

Marczynski, T. J., 1998. Gabaergic deafferentation hypothesis of brain aging and Alzheimer's disease revisited. *Brain research bulletin*, **45**(4):341–379.

Mecocci, P., MacGarvey, U., and Beal, M. F., 1994. Oxidative damage to mitochondrial DNA is increased in Alzheimer's disease. *Annals of neurology*, **36**(5):747–751.

Millon, R., Nicora, F., Muller, D., Eber, M., Klein-Soyer, C., and Abecassis, J., 1989. Modulation of human breast cancer cell adhesion by estrogens and antiestrogens. *Clin Exp Metastasis*, **7**(4):405–15.

Naito, A. T., Akazawa, H., Takano, H., Minamino, T., Nagai, T., Aburatani, H., and Komuro, I., 2005. Phosphatidylinositol 3-kinase-Akt pathway plays a critical role in early cardiomyogenesis by regulating canonical Wnt signaling. *Circ Res*, **97**(2):144–51.

Nelson, H. D., Humphrey, L. L., Nygren, P., Teutsch, S. M., and Allan, J. D., 2002. Postmenopausal hormone replacement therapy - scientific review. *Jama-journal of the American Medical Association*, **288**(7):872–881.

Novaro, V., Roskelley, C. D., and Bissell, M. J., 2003. Collagen-IV and laminin-1 regulate estrogen receptor alpha expression and function in mouse mammary epithelial cells. *Journal of Cell Science*, **116**(Pt 14):2975–86.

Parker, W. D., Filley, C. M., and Parks, J. K., 1990. Cytochrome oxidase deficiency in Alzheimer's disease. *Neurology*, **40**(8):1302–1302.

Parker, W. D., Parks, J., Filley, C. M., and Kleinschmidt-DeMasters, B., 1994. Electron transport chain defects in Alzheimer's disease brain. *Neurology*, **44**(6):1090–1090.

Peterziel, H., Mink, S., Schonert, A., Becker, M., Klocker, H., and Cato, A. C., 1999. Rapid signalling by androgen receptor in prostate cancer cells. *Oncogene*, **18**(46):6322–9.

Weiss, N. S., Ure, C. L., Ballard, J. H., Williams, A. R., and Daling, J. R., 1980. Decreased risk of fractures of the hip and lower forearm with post-menopausal use of estrogen. *New England Journal of Medicine*, **303**(21):1195–1198.

Wong, C. W., McNally, C., Nickbarg, E., Komm, B. S., and Cheskis, B. J., 2002. Estrogen receptor-interacting protein that modulates its nongenomic activity-crosstalk with Src/Erk phosphorylation cascade. *Proc Natl Acad Sci U S A*, **99**(23):14783–8.

Zhang, W. and Liu, H. T., 2002. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res*, **12**(1):9–18.

Zhu, X., Raina, A. K., Lee, H.-g., Casadesus, G., Smith, M. A., and Perry, G., 2004. Oxidative stress signalling in Alzheimer's disease. *Brain research*, **1000**(1):32–39.

Ziel, H. K., 1982. Estrogens role in endometrial cancer. *Obstetrics and Gynecology*, **60**(4):509–515.