BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–9

A Novel Signaling Pathway Impact Analysis (SPIA)

Adi Laurentiu Tarca[1,2], Sorin Draghici[1,3], Purvesh Khatri[1], Sonia S. Hassan[2], Pooja Mittal[2], Jung-sun Kim[2], Chong Jai Kim[2], Juan Pedro Kusanovic[2] & Roberto Romero[2]

[1]Dept. of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202
[2]Perinatology Research Branch-NIH/NICHD, 4 Brush, 3990 John R, Detroit, MI 48201
[3]Corresponding author

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Gene expression class comparison studies may identify hundreds or thousands of genes as differentially expressed (DE) between sample groups. Gaining biological insight from the result of such experiments can be approached, for instance, by identifying the signaling pathways impacted by the observed changes. Most of the existing pathway analysis methods focus on either the number of DE genes observed in a given pathway (enrichment analysis methods), or on the correlation between the pathway genes and the class of the samples (functional class scoring methods). Both approaches treat the pathways as simple sets of genes, disregarding the complex gene interactions that these pathways are built to describe.

Results: We describe a novel Signaling Pathway Impact Analysis (SPIA) that combines the evidence obtained from the classical enrichment analysis with a novel type of evidence, which measures the actual perturbation on a given pathway under a given condition. A bootstrap procedure is used to assess the significance of the observed total pathway perturbation. Using simulations we show that the evidence derived from perturbations is independent of the pathway enrichment evidence. This allows us to calculate a global pathway significance p-value, which combines the enrichment and perturbation p-values. We illustrate the capabilities of the novel method on 4 real data sets. The results obtained on these data show that SPIA has better specificity and more sensitivity than several widely used pathway analysis methods.

Availability: SPIA was implemented as an R package which is available at `http://146.9.88.18/SPIA`.

Contact: sorin@wayne.edu

1 INTRODUCTION

The typical result of a microarray experiment comparing two groups of samples (e.g. normal and diseased) is a list of differentially expressed (DE) genes together with their estimated expression changes between the groups. Translating such results into a better understanding of the underlying biological phenomenon is key to translating the now abundant high-throughput expression data into biological knowledge. An automated approach to map the list of DE genes onto Gene Ontology (GO) terms is the most widely used attempt at this (Khatri *et al.*, 2002; Drăghici *et al.*, 2003; Khatri & Draghici, 2005). More recently, biological annotations have started to include descriptions of gene interactions in the form of gene signaling networks such as KEGG (Ogata *et al.*, 1999), BioCarta (`www.biocarta.com`), and Reactome (Joshi-Tope *et al.*, 2005). This richer type of annotations have opened the possibility of an automatic analysis aimed to identify the gene signaling networks that are relevant in a given condition, and perhaps even the specific signals or signal perturbations involved. This analysis is usually referred to as a pathway analysis. A PubMed search for "microarrays and pathway analysis" returned more than 1800 results, illustrating the numerous attempts to use and develop such techniques. Currently, the pathway analysis with microarray data is primarily performed using the classical approaches inherited from the ontological profiling: over-representation analysis (ORA) (Khatri *et al.*, 2002; Drăghici *et al.*, 2003) and functional class scoring (FCS) (Pavlidis *et al.*, 2004; Goeman *et al.*, 2004; Mootha *et al.*, 2003; Subramanian *et al.*, 2005; Tian *et al.*, 2005). However, both ORA and FCS techniques are limited by the fact that each functional category is analyzed independently without a unifying analysis at a pathway or system level (Tian *et al.*, 2005). This approach is not well suited for a systems biology approach that aims to account for system level dependencies and interactions, as well as identify perturbations and modifications at the pathway or organism level (Stelling, 2004).

Most of the approaches currently available for the analysis of gene signaling networks share a number of important limitations. Firstly, these approaches consider only the set of genes on any given pathway and ignore their position in those pathways. This may be unsatisfactory from a biological point of view. If a pathway is triggered by a single gene product or activated through a single receptor and if that particular protein is not produced, the pathway will be greatly impacted, probably completely shut off. A good example is the insulin pathway (`www.genome.ac.jp/KEGG/pathway/hsa/hsa04910.html`). If the insulin receptor (*INSR*) is not present, the entire pathway is shut off. Conversely, if several genes are involved in a pathway but they only appear somewhere downstream, changes in their expression levels may not affect the given pathway as much. Secondly, some genes have multiple functions and are involved in several pathways but with different roles. For instance, the above *INSR* is also involved in the adherens junction pathway as one of the many receptor protein tyrosine kinases. However, if the expression of *INSR* changes, this

pathway is not likely to be heavily perturbed because *INSR* is just one of many receptors on this pathway. All these aspects are not considered by any of the existing approaches aiming at assessing the impact of a condition on a given signaling pathway. There is a very recent technique (Efroni *et al.*, 2007) however, which takes into account some topological information but this technique aims at phenotype prediction rather than the assessment of given condition which is our primary goal here. Thirdly, and probably the most important current limitation is that the knowledge embedded in these pathways about how various genes interact with each other is largely unexploited. The very purpose of these pathway diagrams is to capture our current knowledge of how genes interact and regulate each other on various pathways. However, the existing analysis approaches consider only the sets of genes involved on these pathways, without taking into consideration their topology. Our understanding of various pathways is expected to improve as more data is gathered. Pathways will be modified by adding, removing or re-directing links on the pathway diagrams. Most existing techniques are completely unable to even sense such changes. Thus, these techniques will provide identical results as long as the pathway diagram involves the same genes, even if the interactions between them are completely re-defined over time. Finally, until now, the expression changes measured in these high throughput experiments have been used only to identify pathways with unexpectedly high number of differentially expressed genes (ORA approaches) or pathways whose genes are clustered in the ranked list of DE genes (FCS methods), but not to directly estimate the impact of such changes on specific pathways. This is also an important limitation. For instance, ORA techniques will see no difference between a situation in which a subset of genes is differentially expressed just above the detection threshold (e.g., 2 fold) and the situation in which the same genes are changing by many orders of magnitude (e.g., 100 fold). Similarly, FCS techniques can provide the same rankings for entire ranges of expression values, if the correlations between the genes and the phenotypes remain similar. Even though analyzing this type of information in a pathway and system context would be extremely meaningful from a biological perspective, currently there is no technique or tool able to do this.

This paper describes a radically different approach that attempts to capture all aspects above. A global probability value, P_G, is calculated for each pathway, incorporating parameters such as the normalized fold-change of the DE genes, the statistical significance of the set of pathway genes, and the topology of the signaling pathway. We recently proposed a technique that combines the pathway topology with the over-representation evidence with very good results (Draghici *et al.*, 2007). However, in this analysis, the evidence measure captured from the pathway topology was not completely independent from the over-representation evidence. In turn, this made the statistic used to rank the pathways more sensitive to noise in the expression data putting too much emphasis on the magnitude of changes. Also, the false positive rates of this method was higher than expected by chance for short lists of DE genes. The approach described here remedies these weaknesses, while retaining the very novel capability of incorporating the pathway topology. The capabilities of the proposed impact analysis are illustrated on a number of real data sets and simulations. We also show that in this technique, the two types of evidence considered are indeed completely independent.

2 SYSTEM AND METHODS

The impact analysis combines two types of evidence: i) the over-representation of DE genes in a given pathway and ii) the abnormal perturbation of that pathway, as measured by propagating measured expression changes across the pathway topology. These two aspects are captured by two independent probability values, P_{NDE} and P_{PERT}.

The first probability, $P_{NDE} = P(X \geq N_{de} \mid H_0)$, captures the significance of the given pathway P_i as provided by an over-representation analysis of the number of DE genes (N_{DE}) observed on the pathway. In the equation above, H_0 stands for the null hypothesis, that the genes that appear as DE on a given pathway are completely random. From a biological perspective this would mean that that the pathway is not relevant to the condition under study. The P_{NDE} value represents the probability of obtaining a number of DE genes on the given pathway at least as large as the observed one, NDE. These P_{NDE} values were obtained assuming that N_{DE} (the number of DE genes on the pathway analyzed) follows a hypergeometric distribution with 3 parameters: m - the number of all pathway genes present on the array, n - the number of genes on the array not belonging to the pathway, k - total number of DE genes. Any of the existing ORA or FCS approaches can be used to calculate P_{NDE}, as long as this probability remains independent of the magnitudes of the fold changes.

The second probability, P_{PERT}, is calculated based on the amount of perturbation measured in each pathway. We define a gene perturbation factor as:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)} \qquad (1)$$

In Eq. (1), the term $\Delta E(g_i)$ represents the signed normalized measured expression change of the gene g_i (log fold-change if two conditions are compared). The second term in Eq.(1) is the sum of perturbation factors of the genes g_j directly upstream of the target gene g_i, normalized by the number of downstream genes of each such gene $N_{ds}(g_j)$. The absolute value of β_{ij} quantifies the strength of the interaction between genes g_j and g_i. These weights have been introduced in order to allow the model to capture the properties of various types of relationships. The results presented in this paper are obtained using all $|\beta| = 1$ in order to minimize the number of model parameters. The sign of β reflects the type of interaction: +1 for induction (activation), -1 for repression and inhibition, as described by each pathway. Note that β will have non-zero value only for the genes that directly interact with the gene g_i according to the pathway description. The work described here used human signaling pathways from KEGG (Ogata *et al.*, 1999). These pathways contain nodes, representing genes/proteins, and directed edges, representing gene signals or interactions such as activation or repression. Given an edge directed from gene/protein A to gene/protein B, we say A is upstream of B, or B is downstream of A.

Equation (1) essentially describes the perturbation factor PF for a gene g_i as a linear function of the perturbation factors of all genes in a given pathway. In the stable state of the system, all relationships must hold, so the set of all equations defining the impact factors for all genes form a system of simultaneous equations whose solution will provide the values for the gene perturbation factors PF_{g_i} (details are provided in the Supplementary material). Subsequently, we calculate the net perturbation accumulation at the

level of each gene, Acc_g, as the difference between the perturbation factor PF of a gene and its observed log fold-change:

$$Acc(g_i) = PF(g_i) - \Delta E(g_i) \tag{2}$$

This subtraction is needed to ensure that DE genes not connected with any other genes will not contribute to the second type of evidence since such genes are already taken into consideration in the over-representation analysis captured by the first term of Eq. 1. In can be shown (see Supplementary material) that the vector of perturbation accumulations Acc can be obtained using the matrix equation:

$$Acc = B \cdot (I - B)^{-1} \cdot \Delta E \tag{3}$$

where B represents the normalized weighted directed adjacency matrix of the graph describing the gene signaling network:

$$B = \begin{pmatrix} \frac{\beta_{11}}{N_{ds(g_1)}} & \frac{\beta_{12}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{1n}}{N_{ds(g_n)}} \\ \frac{\beta_{21}}{N_{ds(g_1)}} & \frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{2n}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\beta_{n1}}{N_{ds(g_1)}} & \frac{\beta_{n2}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix} \tag{4}$$

I is the identity matrix, and

$$\Delta E = \begin{pmatrix} \Delta E(g_1) \\ \Delta E(g_2) \\ \cdots \\ \Delta E(g_n) \end{pmatrix} \tag{5}$$

Only the pathways with non-null determinant of $I - B$ matrix were considered for analysis, even though simple, yet reasonable, transformations of B can be performed to avoid such singularities. Out of the 64 human gene signaling pathways available in KEGG, the majority (52 pathways) satisfy this requirement without any other transformations. The situations in which pathways yield a singular matrix and how these situations can be addressed will be described elsewhere. The total net accumulated perturbation in the pathway is computed as $t_A = \sum_i Acc(g_i)$. The second probability, P_{PERT}, will be the probability to observe a total accumulated perturbation of the pathway, T_A, more extreme than t_A just by chance:

$$P_{PERT} = P(T_A \geq t_A \mid H_0) \tag{6}$$

This probability can be calculated using a bootstrap approach. In this procedure, the same number of DE genes as the one observed on the pathway are allowed to occupy any position in the pathway (random gene IDs) and have any possible log fold-change in the range of those considered by the experimenter to be DE. This allows empirical determination of the null distribution of T_A values (details of the bootstrap procedure are given in the Supplementary materials). Figure 1 illustrates the computation of P_{PERT} for a simple 6 gene pathway containing 2 DE genes. Unlike the classical over-representation approach, the perturbation evidence is shown to be able to capture the importance of the position of the DE genes in the pathway as well as their fold-changes.

The two types of evidence, P_{NDE} and P_{PERT}, are finally combined into one global probability value, P_G, that is used to rank the pathways and test the research hypothesis that the pathway is significantly perturbed in the condition under the study. When the

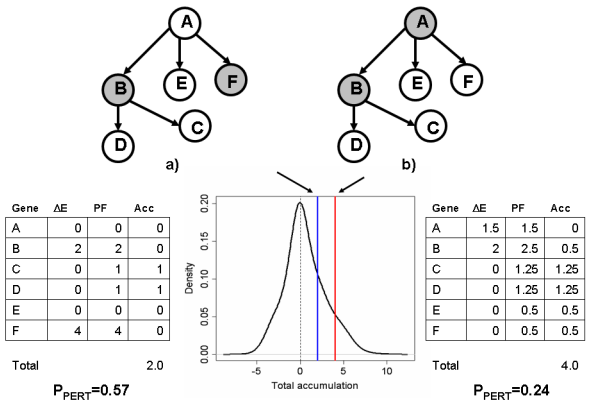

Fig. 1. Capturing the topology of the pathways and the position of the gene through the perturbation analysis. The figure shows a 6-gene pathway with 2 DE genes (shown in grey) in two different situations. One of the two DE genes is in common (gene B) while the second gene is either a leaf node (top left - a), or the entry point in the pathway (top right - b). In a), gene (F) can not perturb the activity of other genes; in b) gene (A) has the ability to influence the activity of all the remaining genes in the pathway, as the topology of the pathway indicates. An over-representation analysis (ORA) would find the two situations equally (in)significant ($P_{NDE} = 0.48$ for a set of 20 monitored genes, out of which 5 are found to be DE). The perturbation evidence extracted by SPIA will give more significance to the situation in b) ($P_{PERT} = 0.24$), even though fold-changes in b) are almost twice as small as those in a) ($P_{PERT} = 0.57$).

null hypothesis is true, the probability of observing a pair of p-values whose product is at least as extreme (low) as the one observed for a given pathway i, $c_i = P_{NDE}(i) \cdot P_{PERT}(i)$, can be shown to be (see Supplementary materials):

$$P_G = c_i - c_i \cdot \ln(c_i) \tag{7}$$

Both components combined within P_G, P_{NDE} and P_{PERT}, are independent of the size of the pathways. P_{NDE} is the probability of observing the given number of DE genes or higher, just by chance. The number of genes expected by chance will increase with the size of the pathway, much like the number of black balls extracted from an urn containing black and white balls will increase with the number of balls extracted in a given trial. Hence, P_{NDE} will be independent of the size of the pathway, much like the hypergeometric probability of extracting a given number of black balls from the urn will automatically take into consideration the number of balls extracted in that particular trial. The second component, P_{PERT} is calculated in a bootstrapping process in which both the pathway and the number of DE genes per pathway are fixed. P_{PERT} will become significant only if the observed fold changes in the observed pathway nodes yield a significantly different impact compared with what is observed on the same pathway when the same number of fake DE genes are thrown in random locations throughout the same pathway. Again, this bootstrap is calculated for each pathway and hence will be independent of the pathway size.

Since P_G is a combined probability value, it can be used not only to rank the pathways, but also to choose a desired level of Type I error. When several tens of pathways are tested simultaneously, as is the case throughout this study, small P_G values can occur also by chance. Therefore we suggest controlling the False Discovery Rate (FDR) of the pathway analysis at 5% by applying the popular FDR algorithm (Benjamini & Yekutieli, 2001).

3 RESULTS AND DISCUSSION

3.1 Absence of false positives under the null hypothesis

From the specificity perspective, an ideal pathway analysis method should not find any significant pathway when a set of randomly selected genes from the reference array are assigned random log fold-changes, regardless of what type of distribution they are drawn from. However, even if the data is completely random (i.e., the null hypothesis is true), any statistical test will reject the null hypothesis for a number of cases directly controlled by the significance threshold, α. It is important, however, to verify that a proposed test does not provide any false positives beyond this expected proportion. In order to verify that SPIA does not provide a number of false positives above the significance threshold, we performed a number of simulations of the null hypothesis that can be divided into 3 scenarios. A reasonable scenario in which one should not find significant pathways is when the differentially expressed (DE) genes have random normal log fold-changes, and the genes are selected at random. In this setup (further referred to as scenario I), we select N_{de} random genes as DE from a reference array of size 20,000. The reference array includes all genes from all 52 pathways analyzed. The genes were assigned log fold-changes from a random normal distribution, $N(0,1)$. This is illustrated in top left panel of Fig. 2. An alternative model for the null hypothesis is an experiment in which one compares two groups of samples among which there are no real biological differences. However, due to various reasons such as improper normalization or array batch effects problems, the values measured for various genes will be different. Thus, one can always falsely identify some genes as DE (using for instance a fold-change selection method (Drăghici, 2002)). In this case, the distribution of the log fold-changes will be bimodal (scenario II). This is illustrated in the top center panel of Fig. 2. In this second scenario, log fold-changes are still drawn from a random normal distribution, $N(0,1)$ but they are restricted to be at least one standard deviation away from the mean. Another particularly interesting situation is when all DE genes log-fold changes are either positive or negative (all genes are up regulated or down regulated). This situation is illustrated in the top right panel of Fig. 2. In this case the log fold-changes of the so called DE genes may have a unimodal distribution but they will be all far from 0 and share the same sign, e.g., a random normal distribution with mean 3 and standard deviation of 0.5 (scenario III). It should be noted that, from this perspective, the main limitation of the classical hypergeometric analysis turns into an advantage: since the hypergeometric enrichment analysis does not take into consideration either the specific fold-changes, nor the pathway topology, it will not be susceptible to false positives due to such causes as described above. Indeed, this is illustrated in the middle panel of Fig. 2 which shows that the distribution of the hypergeometric p-value is essentially uniform in all 3 scenarios considered. The results presented in Fig. 2 show that SPIA also yields a uniform distribution of p-values under the null hypothesis and therefore will provide no false positives beyond the unavoidable level equal to the chosen significance threshold, α. This is true regardless of the number of DE genes analyzed (Table 1).

Table 1. The false positive rates and the correlation coefficients between P_{PERT} and P_{NDE}, as a function of the number of DE genes analyzed averaged over the three scenarios depicted in Fig. 2. The data shows no correlation between P_{PERT} and P_{NDE}, as well as an average false positive rate at the expected 5% level.

N_{de}	FP(ORA)	FP(SPIA)	R^2
100	0.072	0.068	0.0022
300	0.045	0.048	0.0028
500	0.038	0.044	0.0032
1000	0.036	0.045	0.0041
2000	0.038	0.045	0.0041
5000	0.036	0.046	0.0002

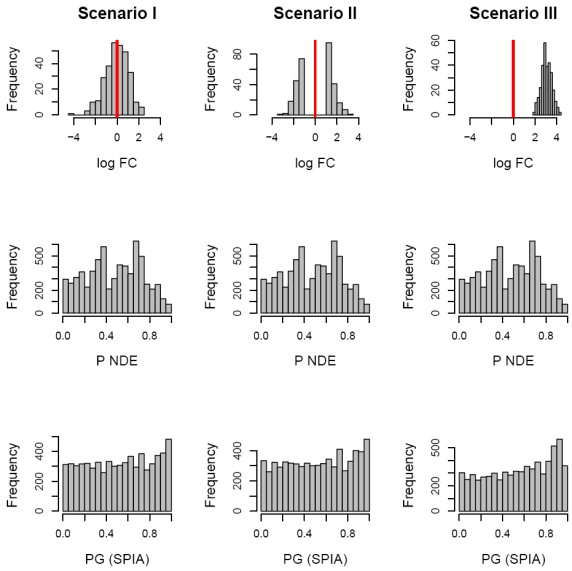

Fig. 2. Distribution of p-values under 3 null distribution scenarios for the hypergeometric and SPIA models. $N_{de} = 300$ gene IDs were selected at random out of 20,000 possible IDs containing all genes on all 52 pathways analyzed. The randomly selected gene IDs were assigned log fold-changes from i) a random normal distribution $N(0,1)$; ii) a bimodal distribution obtained by sampling from the tails of a $N(0,1)$ distribution; and iii) random normal $N(3,0.5)$. For each scenario, the experiment was repeated 200 times and P_{NDE}, P_{PERT} and P_G were computed for all pathways receiving at least one DE gene. The resulting p-values for all pathways and all iterations were pooled together and shown as histograms for ORA (P_{NDE}) and SPIA (P_G) on rows 2 and 3 respectively. The false positives rates for SPIA at $\alpha = 5\%$ were 4.7%, 5.0% and 4.6%, in scenarios I, II and III, respectively. For ORA, the same positive rates were 4.5% in all 3 scenarios. False positive rates as an average over these 3 scenarios are provided in Table 1 for several values of N_{de}.

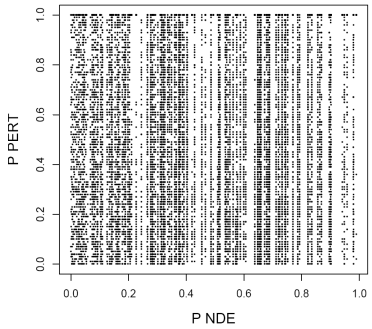

Fig. 3. Correlation analysis between P_{NDE} and P_{PERT} under the null hypothesis. This scatter-plot shows all pairs of p-values for 52 pathways, 200 random trials, and the 3 fold-change distribution scenarios considered. As shown in Table 1, the squared correlation coefficient, R^2, was less than 0.005, regardless of the number of genes analyzed, N_{de}. The current plot was obtained with $N_{de} = 300$.

3.2 The independence of the perturbation and enrichment statistics

The fact that P_{PERT} and P_{NDE} are two independent variables under the null hypothesis, is theoretically justified by the design of the bootstrap procedure used to compute P_{PERT}. This independence has also been verified using a simulation, as follows. A set of $N_{de} = 300$ gene IDs was selected at random out of 20,000 possible IDs containing all genes on all 52 pathways analyzed. The randomly selected gene IDs were assigned log fold-changes from i) a random normal distribution $N(0, 1)$; ii) a bimodal distribution obtained by sampling from the tails of a $N(0, 1)$ distribution; and iii) random normal $N(3, 0.5)$. For each scenario, the experiment was repeated 200 times and P_{NDE}, P_{PERT} were computed for all pathways receiving at least one DE gene. The resulting pairs of P_{NDE} and P_{PERT} are shown as a scatter plot in Fig. 3. The squared correlation coefficient was $R^2 = 0.0028$, indicating essentially no correlation between the two p-values. This remains true regardless of the number of DE genes analyzed (see Table 1). These simulations prove that the evidence from perturbations as computed by SPIA is linearly independent from the over-representation evidence under the null hypothesis.

3.3 Sensitivity and pathways ranking on real data sets

Assessing the capabilities of any pathway analysis method in real experiments is a challenge in itself because the ground truth is never known. In the absence of a gold standard, the best alternatives are to: i) analyze the results of the pathway analysis method in the light of the existing biological knowledge regarding the condition studied, and ii) compare it with the existing methods in the context of the same existing biological knowledge. The absence of a definitive answer regarding the involvement of a given pathway in a given condition makes it impossible to calculate exact values for sensitivity, specificity, ROCs, etc. However, the methods can be compared in terms of the number of pathways that are found to be significant in a given condition and how well the significant pathways fit with the existing biological knowledge. This type of assessment is the current best practice in this area (Subramanian *et al.*, 2005).

We used four data sets in order to illustrate the capabilities of the newly proposed pathway analysis method, SPIA. The first such dataset compares 12 colorectal cancer samples with 10 normal samples (Hong *et al.*, 2007) using Affymetrix HG-U133 Plus 2.0. microarray platform. This dataset is available via the Gene Expression Omnibus (ID=GSE4107) and it will be referred to as the *Colorectal cancer* dataset. Several pathways are known to be relevant to the colorectal cancer, including the *Colorectal pathway* itself, the *PPAR signaling pathway* (Shureiqi *et al.*, 2003) and *MAPK signaling pathway* (Fang & Richardson, 2005).

The second data set is the result of comparing gene expression levels in cervix tissue samples from women at term with spontaneous labor (TL group) ($n = 9$) and those at term without labor (TNL group) ($n = 7$). The microarray platform used was Affymetrix HG-U133 Plus 2.0. The details of this study and its biological significance are described elsewhere (Hassan *et al.*, 2006, 2007). This data set will be referred to as *LaborC*.

The third data set is the result of comparing gene expression levels between umbilical veins (UV)(n=6) and umbilical arteries (UA) tissues (n=6) using Illumina BeadChips Human-6 V2 arrays. This data set will be referred to as *Vessels* data set. The details of this study and its biological significance are presented elsewhere (Kim *et al.*, 2008). This data set is available in the Arrays Express repository (ID: E-TABM-368).

The fourth data set used in this study was produced by comparing gene expression levels in myometrium tissue samples from women at term with spontaneous labor ($n = 27$) and those at term without labor ($n = 30$) using Affymetrix HG-U133 Plus 2.0 platform. In essence, this experiment studies the same medical condition as the second data set (spontaneous labor at term) except that the investigated tissue is different (myometrium rather than cervix) and the numbers of samples are much larger for each class (27/30 rather than 9/7). The details of this study and its biological significance are not published yet. This data set will be referred to as *LaborM*.

After proper preprocessing, including $\log_2$ transformation and quantile normalization (Irizarry *et al.*, 2003), microarray expression data from all three experiments were analyzed in the same way. Differential expression was inferred using a moderated t-test (Smyth, 2005) and a false discovery rate (FDR) adjustment of the resulting p-values. For the Colorectal cancer and Vessels datasets, genes were considered as DE provided they had a FDR corrected p-value less than 0.05. For the LaborC and LaborM data sets, genes were considered as DE provided they had a FDR corrected p-value less than 0.05 and 0.01, respectively, and their fold-change was greater than 2 and 1.5, respectively. The additional stringency used on the gene selection for these later two datasets was implemented in order to be consistent with the original analyzes of the authors.

The richness and suitability of this ensemble of four datasets can be discussed from three different perspectives. Firstly, they were obtained using two different microarray platforms (Affymetrix and Illumina). Secondly, the distribution of log fold-changes for the DE genes has different properties among the data sets, since for the Colorectal cancer and Vessels data no threshold on fold-changes was used, as opposed to the other two datasets. Finally, for all datasets there are several biological clues about what pathways are expected to be involved. Also, since both LaborC and LaborM datasets studied the impact of spontaneous labor in two closely

related uterine regions, a number of similarities can be expected between the pathways impacted in these two experiments.

The SPIA algorithm was compared to several existing pathway analysis methods including the classical over-representation analysis (ORA) using a hypergeometric model, and the Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005). The comparison was based on statistical power (the ability to find significant pathways), specificity (the ability to limit the number of false positive pathways), as well as the ability to provide a meaningful ranking of the pathways analyzed.

A significance threshold of 5% was used on the False Discovery Rate (FDR) corrected p-values in order to infer pathway significance. For both SPIA and ORA the FDR adjusted p-values were computed from the nominal p-values using the R function "p.adjust", while for GSEA, the FDR values (also called q-values) are reported as provided by the R GSEA V 1.0. Only the top 15 pathways are given in tables for each analysis method, with pathway names being truncated to save space. The correspondence between the pathway IDs shown in these tables and the full KEGG pathway names can be found at `ftp://ftp.genome.jp/pub/kegg/xml/organisms/hsa/index.html`.

On the Colorectal cancer dataset, the three pathway analysis methods were compared in terms of their ability to identify the Colorectal cancer pathway, the PPAR signaling pathway (Shureiqi *et al.*, 2003) and MAPK signaling pathway (Fang & Richardson, 2005) as relevant to colorectal cancer disease. Using a 5% cut-off of the FDR adjusted p-values, both ORA and SPIA identified the PPAR signaling pathway and MAPK signaling pathway as significant to the condition under the study (see Tables 3 and 2). However, only SPIA identified the Colorectal cancer pathway itself as significant (see Table 2). This was possible due to additional evidence from perturbations (PPERT=0.04 in Table 2). In addition, SPIA downgraded the *Alzheimers disease pathway* from second position in the top with ORA to the fourth position. This pathway is most likely not relevant to Colorectal cancer, and it appears among the significant pathways for both ORA and SPIA because 14 genes of this pathway are DE, out of all 22 genes of this pathway that are represented on the reference array. On the other hand GSEA identified no significant pathway on this dataset (see Tables 1 and 2 in Supplementary Material). Also, according to GSEA, the top ranked pathways for this dataset are the *Huntington's disease* and *Parkinson's disease* pathways, which are not likely to be relevant to colorectal cancer.

For the LaborC dataset, SPIA identified *Cytokine-cytokine receptor interaction*, *Complement and coagulation cascades*, *Focal adhesion* and *ECM-receptor interaction* pathways as being the most significantly impacted pathways in this condition (Table 4). These pathways are also the top 4 found by ORA. SPIA clearly indicates that there are no other significant pathways beyond these top four (the next most significant p-value is 0.46 for the FWER correction). This difference of over two orders of magnitude is independent of the type of correction (no correction, FDR or FWER) and represents a clear demarcation, also independent of the choice of any of the usual significance threshold: 1%, 5%. In contrast, the classical ORA results place the pathways in a continuum of p-values in which the choice of the multiple correction method and that of the significance threshold can significantly change the results (Table 5). For instance, two additional pathway are reported as significant at the usual 5% significance on the FDR corrected values, while at

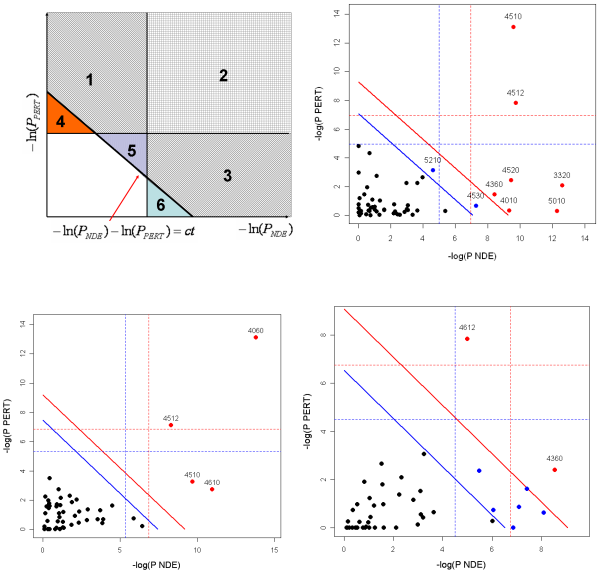

Fig. 4. Two-dimensional plots illustrating the relationship between the two types of evidence considered by SPIA. The X-axis shows the over-representation evidence while the Y-axis shows the perturbation evidence. In the top-left plot, areas 2, 3 and 6 together will include pathways that meet the over-representation criterion ($P_{NDE} < \alpha$). Areas 1, 2 and 4 together will include pathways that meet the perturbation criterion ($P_{PERT} < \alpha$). Areas 1, 2, 3 and 5 will include the pathways that meet the combined SPIA criteria ($P_G < \alpha$). Note how SPIA results are different from a mere logical operation between the two criteria (OR would be areas 1, 2, 3, 4 and 6; AND would be area 2). Interestingly, SPIA removes those pathways that are supported by evidence of any one single type that is just above their corresponding thresholds but not supported by the other type of evidence (areas 4 and 6), but adds pathways that are just under the individual significance thresholds but supported by both types of evidence (area 5). The other plots show the pathway analysis results on the *Colorectal cancer* (top right), *LaborC* (bottom left) and *Vessels* (bottom right) datasets. Each pathway is represented by a point. Pathways above the oblique red line are significant at 5% after Bonferroni correction, while those above the oblique blue line are significant at 5% after FDR correction. The vertical and horizontal thresholds represent the same corrections for the two types of evidence considered individually. Note that for the *Colorectal cancer* dataset (top right), the Colorectal cancer pathway (ID=5210) is only significant according to the combined evidence but not so according to any individual evidence P_{NDE} or P_{PERT}.

the 10%, there are three such additional pathway. Unfortunately, the most "significant" of these additional pathways is the *Renal cell carcinoma* pathway, which in fact is very unlikely to be truly relevant to pregnancy and labor.

On the other hand, GSEA was unable to find any significant pathways in this condition (after either FDR or FWER) suggesting a more limited power (Table 6 and 7). Furthermore, GSEA produces two tables, one for the term labor (TL), which is the normal condition in this case, and one for the non-labor at term (TNL). In general, it is unclear what the biological interpretation would be for something which is found to be significantly perturbed in the normal condition (this is not the case here because GSEA does not find any pathway to be significant in this condition).

Table 2. SPIA results on Colorectal cancer dataset; [1,2]Calculated for P_G; Act.=Activated, Inh.=Inhibited

KEGG Pathway	P_{NDE}	P_{PERT}	P_G	$P_{FDR}{}^1$	$P_{FWER}{}^2$	Status
Focal adhe..4510	0.0001	0.0000	0.0000	0.00000	0.00000	Act.
ECM-recept..4512	0.0001	0.0004	0.0000	0.00001	0.00002	Act.
PPAR signa..3320	0.0000	0.1240	0.0000	0.00011	0.00034	Inh.
Alzheimers..5010	0.0000	0.7260	0.0001	0.00059	0.00235	Act.
Adherens j..4520	0.0001	0.0852	0.0001	0.00090	0.00452	Act.
Axon guida..4360	0.0002	0.2324	0.0006	0.00487	0.02922	Act.
MAPK signa..4010	0.0001	0.7112	0.0007	0.00504	0.03527	Inh.
Tight junc..4530	0.0007	0.5156	0.0032	0.02073	0.16585	Act.
Colorectal..5210	0.0098	0.0432	0.0037	0.02151	0.19359	Act.
Wnt signal..4310	0.0187	0.0704	0.0101	0.05227	0.52268	Inh.
Renal cell..5211	0.0259	0.1048	0.0188	0.08869	0.97561	Inh.
Regulation..4810	0.0046	0.7328	0.0226	0.09797	1.00000	Act.
Thyroid ca..5216	0.0516	0.1032	0.0332	0.13284	1.00000	Inh.
Cytokine-c..4060	0.5049	0.0132	0.0401	0.14880	1.00000	Act.
Antigen pr..4612	0.9948	0.0080	0.0464	0.16094	1.00000	Act.

Table 3. ORA results on Colorectal cancer dataset; [3,4]Calculated for P_{NDE}.

KEGG Pathway	P_{NDE}	$P_{FDR}{}^3$	$P_{FWER}{}^4$
PPAR signa..3320	0.00000	0.00012	0.00017
Alzheimers..5010	0.00000	0.00012	0.00024
ECM-recept..4512	0.00006	0.00078	0.00308
Focal adhe..4510	0.00007	0.00078	0.00360
Adherens j..4520	0.00008	0.00078	0.00411
MAPK signa..4010	0.00009	0.00078	0.00465
Axon guida..4360	0.00022	0.00165	0.01156
Tight junc..4530	0.00069	0.00450	0.03599
Regulation..4810	0.00461	0.02665	0.23982
Colorectal..5210	0.00983	0.05113	0.51134
Wnt signal..4310	0.01871	0.08843	0.97275
Renal cell..5211	0.02591	0.10365	1.00000
Complement..4610	0.02591	0.10365	1.00000
Insulin si..4910	0.03546	0.13172	1.00000
Gap juncti..4540	0.04630	0.14911	1.00000

Table 4. SPIA results on the LaborC dataset. [1,2]Calculated for P_G; Act.=Activated, Inh.=Inhibited

KEGG Pathway	P_{NDE}	P_{PERT}	P_G	$P_{FDR}{}^1$	$P_{FWER}{}^2$	Status
Cytokine-c..4060	0.0000	0.0000	0.0000	0.0000	0.0000	Act.
ECM-recept..4512	0.0002	0.0008	0.0000	0.0001	0.0002	Act.
Complement..4610	0.0000	0.0652	0.0000	0.0003	0.0008	Inh.
Focal adhe..4510	0.0001	0.0384	0.0000	0.0004	0.0016	Act.
Renal cell..5211	0.0016	0.8032	0.0097	0.0804	0.4652	Act.
Jak-STAT s..4630	0.0028	0.4764	0.0100	0.0804	0.4823	Act.
Phosphatid..4070	0.0111	0.1968	0.0156	0.1067	0.7467	Act.
mTOR signa..4150	0.0238	0.2760	0.0396	0.2378	1.0000	Inh.
Regulation..4810	0.0198	0.5080	0.0563	0.2804	1.0000	Inh.
Type II di..4930	0.0533	0.2568	0.0724	0.2804	1.0000	Inh.
MAPK signa..4010	0.0211	0.6532	0.0729	0.2804	1.0000	Act.
Toll-like ..4620	0.0282	0.5104	0.0754	0.2804	1.0000	Act.
Circadian ..4710	0.1100	0.1320	0.0759	0.2804	1.0000	Inh.
Huntington..5040	0.1716	0.0996	0.0867	0.2971	1.0000	Inh.
Epithelial..5120	0.6281	0.0308	0.0957	0.3061	1.0000	Act.

Table 5. ORA results on the LaborC dataset. [3,4]Calculated for P_{NDE}.

KEGG Pathway	P_{NDE}	$P_{FDR}{}^3$	$P_{FWER}{}^4$
Cytokine-c..4060	0.0000	0.0000	0.0000
Complement..4610	0.0000	0.0004	0.0008
Focal adhe..4510	0.0001	0.0010	0.0030
ECM-recept..4512	0.0002	0.0029	0.0117
Renal cell..5211	0.0016	0.0151	0.0755
Jak-STAT s..4630	0.0028	0.0221	0.1326
Phosphatid..4070	0.0111	0.0760	0.5323
Regulation..4810	0.0198	0.1127	0.9497
MAPK signa..4010	0.0211	0.1127	1.0000
mTOR signa..4150	0.0238	0.1144	1.0000
Toll-like ..4620	0.0282	0.1230	1.0000
Type II di..4930	0.0533	0.2100	1.0000
Insulin si..4910	0.0569	0.2100	1.0000
Cell cycle..4110	0.0923	0.3117	1.0000
TGF-beta s..4350	0.0974	0.3117	1.0000

For the Vessels dataset, it can be argued (Kim *et al.*, 2008) that the main difference between the umbilical veins and arteries is their pro-inflammatory behavior, therefore one of the most biologically meaningful pathways is the *Antigen processing and presentation* (pathway 4612 in Fig. 4, bottom left panel). Indeed, this pathway was identified by SPIA as the most significant (raw $p < 0.00005$, $p = 0.0016$ after either FWER or FDR, see Table 3 in Supplementary materials). In contrast, the classical ORA ranks this pathway only on the 9th place, with a p-value that may or may not be significant depending on the type of correction and significance threshold (p=0.288 after FWER, p=0.0321 after FDR). With the most stringent correction (FWER) and significance threshold (0.01), SPIA identifies one additional pathway in this condition: *Axon guidance*. This is in agreement with the ORA which reports this pathway as the most significant in this condition. Both methods report this because 12 out of the 127 genes on this pathway are DE in this condition, which is about 4 times more than expected by chance. On this data, ORA identified 9 significant pathways, SPIA identified 8, while GSEA none (see Tables 3-6 in the Supplementary material).

The LaborM dataset studied the same medical condition as the LaborC, except that the samples were collected from the myometrium, rather than cervix, both parts of the (same) uterus. It is therefore reasonable to expect that some of the pathways involved in LaborC and LaborM would be common. Both ORA and SPIA ranked the *Cytokine-cytokine receptor interaction* as the most relevant pathway. The roles of cytokines in the human myometrium in labor have been previously assessed in several investigations, indicating their biological significance in human parturition (Breuiller-Fouche & Germain, 2006). Studies on uterine macrophages in the myometrium showed that the inflammatory cytokines, $IL1B$ and $TNFa$, are important regulators of $PGHS2$ and $IL8$ (Tattersall *et al.*, 2008). An increase in $IL1$, $IL6$, $IL8$ and $TNFa$ within tissues of the laboring uterus and cervix is well-known (Osman *et al.*, 2003). Increased mRNA expression of $CCL13$, $CCL19$, $CCL21$, $CXCR4$ and $CXCR5$ in the

Table 6. GSEA results on the LaborC dataset, enrichement in TL group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR (median)	global p-val
Cytokine-c..4060	0.0142	0.58128	0.341	0	0.263
Toll-like ..4620	0.0513	0.51643	0.6175	0.38462	0.179
Jak-STAT s..4630	0.0515	0.60647	0.541	0.45455	0.2435
Complement..4610	0.0665	0.41882	0.646	0.3125	0.106
Adipocytok..4920	0.0858	0.33712	0.709	0.26316	0.0525
Type II di..4930	0.0926	0.32744	0.8105	0.25641	0.0425
ECM-recept..4512	0.1102	0.38869	0.6945	0.3	0.0775
Maturity o..4950	0.1407	0.38816	0.786	0.30612	0.063
Focal adhe..4510	0.1421	0.34761	0.7925	0.27778	0.05
Epithelial..5120	0.1626	0.32478	0.8405	0.26471	0.037
MAPK signa..4010	0.1814	0.43041	0.931	0.375	0.0615
Regulation..4810	0.1818	0.44217	0.96	0.38961	0.0515
Renal cell..5211	0.1998	0.41124	0.9385	0.35503	0.046
Type I dia..4940	0.2083	0.40331	0.912	0.34091	0.056
Colorectal..5210	0.3005	0.57367	0.9925	0.55556	0.082

Table 7. GSEA results on the LaborC dataset, enrichement in TNL group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR (median)	global p-val
Notch sign..4330	0.0468	0.9707	0.5935	0.75	0.382
Tight junc..4530	0.1083	0.5329	0.8295	0.42568	0.1325
Ubiquitin ..4120	0.1104	0.7213	0.707	0.55263	0.2715
Dentatorub..5050	0.1554	0.6512	0.805	0.51852	0.2185
Endometria..5213	0.1845	0.7443	0.9475	0.64615	0.26
Phosphatid..4070	0.1934	0.6644	0.971	0.6	0.195
GnRH signa..4912	0.1946	0.6939	0.9625	0.6087	0.221
Olfactory ..4740	0.3213	0.5848	0.9725	0.525	0.1295
Hedgehog s..4340	0.3350	0.6803	0.9935	0.64412	0.164
Cell cycle..4110	0.3377	0.5825	0.9805	0.52859	0.1205
Amyotrophi..5030	0.3672	0.7403	0.993	0.7	0.223
Basal cell..5217	0.3685	0.6308	0.9935	0.60577	0.117
Thyroid ca..5216	0.4229	0.6429	0.9955	0.62821	0.1085
Gap juncti..4540	0.4621	0.6696	0.9985	0.66422	0.126
Melanogene..4916	0.5304	0.7081	1	0.7	0.132

myometrium has also been shown (Bethin *et al.*, 2003; Breuiller-Fouche & Germain, 2006). The Cytokine-cytokine receptor interaction pathway was also the most relevant in the LaborC dataset according with these two methods. In contrast, GSEA failed to identify this pathway as significant in either LaborM or LaborC sets (Table 6-7 and Supplementary Tables 9-10). This is disappointing in the light of the fact each group had approx. 30 samples, therefore justifying the expectation of a reasonable statistical power. Furthermore, GSEA's rankings of this pathway were not consistent between these two related data sets. Combined with the low false positive rate verified on various null hypothesis simulations, the results on all three datasets suggest a higher statistical power for SPIA when compared to GSEA.

4 CONCLUSIONS

This study introduced the Signaling Pathway Impact Analysis (SPIA) which provides increased sensitivity when compared to GSEA, as well as improved specificity and better pathway ranking when compared to ORA. The SPIA algorithm was implemented as a standard R library available on request from the corresponding author. We also plan to make this library freely available as part of the Bioconductor project (www.bioconductor.org).

ACKNOWLEDGEMENTS

This research was supported, in part, by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. SD was also supported by the following grants: NSF DBI 0234806, CCF 0438970, 1R01HG003491, 1U01CA117478, 1R21CA100740, 1R01 NS045207, 5R21EB000990, 2P30 CA022453. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, NIH, DOD or any other of the funding agencies. The authors declare that they have no competing financial interests.

REFERENCES

Benjamini, Y. & Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics,* **29** (4), 1165–1188.

Bethin, K. E., Nagai, Y., Sladek, R., Asada, M., Sadovsky, Y., Hudson, T. J. & Muglia, L. J. (2003) Microarray analysis of uterine gene expression in mouse and human pregnancy. *Mol Endocrinol,* **17** (8), 1454–69.

Breuiller-Fouche, M. & Germain, G. (2006) Gene and protein expression in the myometrium in pregnancy and labor. *Reproduction,* **131** (5), 837–50.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C. & Romero, R. (2007) A systems biology approach for pathway level analysis. *Genome Research,* **17** (10), 1537–1545.

Drăghici, S. (2002) Statistical intelligence: effective analysis of high-density microarray data. *Drug Discovery Today,* **7** (11), S55–S63.

Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. & Krawetz, S. A. (2003) Global functional profiling of gene expression. *Genomics,* **81** (2), 98–104.

Efroni, S., Schaefer, C. F. & Buetow, K. H. (2007) Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS ONE,* **2** (5), e425.

Fang, J. & Richardson, B. (2005) The mapk signalling pathways and colorectal cancer. *Lancet Oncol,* **6**, 322–327.

Goeman, J. J., van de Geer, S. A., de Kort, F. & van Houwelingen, H. C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics,* **20** (1), 93–99.

Hassan, S. S., Romero, R., Haddad, R., Hendler, I., Khalek, N., Tromp, G., Diamond, M. P., Sorokin, Y. & Malone, J., J. (2006) The transcriptome of the uterine cervix before and after spontaneous term parturition. *Am J Obstet Gynecol,* **195** (3), 778–86.

Hassan, S. S., Romero, R., Tarca, A. L., Draghici, S., Pineles, B., Bugrim, A., Khalek, N., Camacho, N., Mittal, P., Yoon, B. H., Espinoza, J., Kim, C. J., Sorokin, Y. & Malone, J. J. (2007) Signature pathways identified from gene expression profiles in the human uterine cervix before and after spontaneous term parturition. *Am J Obstet Gynecol,* **197** (3), 250.e1–250.e7.

Hong, Y., Ho, K. S., Eu, K. W. & Cheah, P. Y. (2007) A susceptibility gene set for early onset colorectal cancer

that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res,* **13** (4), 1107–14.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. & Speed, T. P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics,* **to appear**.

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., Lewis, S., Birney, E. & Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research,* **33** (Database issue), D428–432.

Khatri, P. & Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics,* **21** (18), 3587–3595.

Khatri, P., Drăghici, S., Ostermeier, G. C. & Krawetz, S. A. (2002) Profiling gene expression using Onto-Express. *Genomics,* **79** (2), 266–270.

Kim, J.-S., Romero, R., Tarca, A. L., LaJeunesse, C., Han, Y. M., Kim, M. J., Suh, Y. L., Draghici, S., Mittal, P., Gotsch, F., Kusanovic, J. P., Hassan, S. & Kim, C. J. (2008) Gene expression profiling demonstrates a novel role for fetal fibrocytes and the umbilical vessels in human fetoplacental development. *The Journal of Cellular and Molecular Medicine,* **to appear**.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. (2003) PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics,* **34** (3), 267–273.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. *et al.* (1999) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* **27** (1), 29–34.

Osman, I., Young, A., Ledingham, M. A., Thomson, A. J., Jordan, F., Greer, I. A. & Norman, J. E. (2003) Leukocyte density and pro-inflammatory cytokine expression in human fetal membranes, decidua, cervix and myometrium before and during labour at term. *Mol Hum Reprod,* **9** (1), 41–5.

Pavlidis, P., Qin, J., Arango, V., Mann, J. J. & Sibille, E. (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research,* **29** (6), 1213 –1222.

Shureiqi, I., Jiang, W., Zuo, X., Wu, Y., Stimmel, J., Leesnitzer, L., Morris, J., Fan, H., Fischer, S. & Lippman, S. (2003) The 15-lipoxygenase-1 product 13-s-hydroxyoctadecadienoic acid down-regulates ppar-delta to induce apoptosis in colorectal cancer cells. *Proc Natl Acad Sci USA,* **100** (17), 9968–73.

Smyth, G. K. (2005) *Limma: linear models for microarray data*. New York: Springer pp. 397–420.

Stelling, J. (2004) Mathematical models in microbial systems biology. *Current opinion in microbiology,* **7** (5), 513–8.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceeding of The National Academy of Sciences of the USA,* **102** (43), 15545–15550.

Tattersall, M., Engineer, N., Khanjani, S., Sooranna, S. R., Roberts, V. H., Grigsby, P. L., Liang, Z., Myatt, L. & Johnson, M. R. (2008) Pro-labour myometrial gene expression: are preterm labour and term labour the same? *Reproduction,* **135** (4), 569–79.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S. & Park, P. J. (2005) Discovering statistically significant pathways in expression profiling studies. *Proceeding of The National Academy of Sciences of the USA,* **102** (38), 13544–13549.

BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–2

Supplementary Material for: A Novel Signaling Pathway Impact Analysis (SPIA)

Adi Laurentiu Tarca[1,2], Sorin Draghici[1,3], Purvesh Khatri[1], Sonia S. Hassan[2], Pooja Mittal[2], Jung-sun Kim[2], Chong Jai Kim[2], Juan Pedro Kusanovic[2] & Roberto Romero[2]

[1]Dept. of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202
[2]Perinatology Research Branch-NIH/NICHD, 4 Brush, 3990 John R, Detroit, MI 48201
[3]Corresponding author

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

1 COMPUTING PERTURBATION FACTORS

Let us consider the normalized weighted directed adjacency matrix of the graph describing the gene signaling network:

$$B = \begin{pmatrix} \frac{\beta_{11}}{N_{ds(g_1)}} & \frac{\beta_{12}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{1n}}{N_{ds(g_n)}} \\ \frac{\beta_{21}}{N_{ds(g_1)}} & \frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{2n}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\beta_{n1}}{N_{ds(g_1)}} & \frac{\beta_{n2}}{N_{ds(g_2)}} & \cdots & \frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix} \quad (1)$$

In this matrix, β_{ij} is the efficiency with which a unit perturbation of gene j is propagated to gene i, and $N_{ds}(g_i)$ is the number of genes downstream of gene g_i. node) would sum up to 1 if taken in absolute values.

Let the vector of measured log fold-changes be:

$$\Delta E = \begin{pmatrix} \Delta E(g_1) \\ \Delta E(g_2) \\ \cdots \\ \Delta E(g_n) \end{pmatrix} \quad (2)$$

If a gene is not differentially expressed, its log fold-change is assigned the value 0. The vector of gene perturbation factors is:

$$PF = \begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} \quad (3)$$

Then, the equations defining the perturbations after reaching a stable state:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)} \quad (4)$$

can be re-written as:

$$PF = \Delta E + B \cdot PF \quad (5)$$

while the net accumulations of the perturbations:

$$Acc(g_i) = PF(g_i) - \Delta E(g_i) \quad (6)$$

can also be re-written as:

$$Acc = PF - \Delta E = B \cdot PF \quad (7)$$

From Eq. 5 and 7, and assuming that the matrix $I-\beta$ is non-singular, we can calculate:

$$Acc = B \cdot (I - B)^{-1} \cdot \Delta E \quad (8)$$

2 BOOTSTRAP PROCEDURE FOR COMPUTING A P-VALUE FROM PATHWAY PERTURBATIONS.

The computation of P_{PERT} for a given pathway is based on a bootstrap procedure in which we want to test if the observed global activation or inhibition of the pathway computed with the real data, t_A is unusual compared to a multitude of random scenarios. The step by step procedure we used is:

1. An iteration counter k is initialized ($k = 1$).
2. A set of $N_{de}(P_i)$ gene IDs is selected at random from the pathway P_i where the $N_{de}(P_i)$ is the number of DE genes observed on the pathway with the real data. The log fold-changes for these random gene IDs are assigned by drawing a random sample with replacement from the distribution of all DE genes to be analyzed. item Eq. 8 is used to compute the perturbation accumulations Acc, for each gene in P_i. The net total accumulation is computed as the sum of all perturbation accumulations across each pathway: $T_A(k) = \sum_i Acc(g_{ik})$.
3. Steps 2 and 3 above are repeated a large number of times ($N_{ite} = 2000$).
4. The median of T_A is computed and subtracted from $T_A(k)$ values centering their distribution around 0. The resulting corrected values are denoted with $T_{A,c}(k)$. The observed net total accumulation is also corrected for the shift in the null distribution median to give, $t_{A,c}$.
5. If $t_{A,c}$ is positive then we conclude that the pathway is activated (or positively perturbed). If $t_{A,c}$ is negative then we assume that the pathway is inhibited (or negatively perturbed).

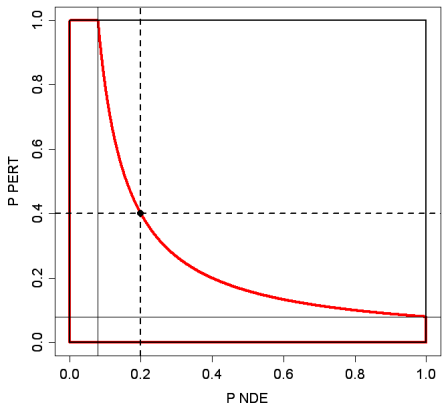

Fig. 1. Combining P_{NDE} and P_{PERT} into a single probability value, P_G. The black rectangle [0,1]x[0,1] contains all possible values that P_{NDE} and P_{PERT} can take. The curve shown is the locus of all combinations of 2 p-values that have the same product $P_{NDE} \cdot P_{PERT} = c$ (which for this example is: $c = 0.2 \cdot 0.4 = 0.08$). The points under and to the left of this curve represent all combinations that would yield a product less than 0.08. The red contour designates the surface whose area is P_G for the chosen example of the pair ($P_{NDE} = 0.2$ and $P_{PERT} = 0.4$) (black dot), under the null hypothesis. The P_G is the probability to have such a combination which can be quantified as the ratio of the area under the curve divided by the entire area of the square (which is 1). In this case, $P_G = 0.282$.

6. The probability to observe such total net inhibition or activation just by chance, P_{PERT}, is computed as:

$$P_{PERT} = \begin{cases} 2 \cdot \frac{\sum_k I(T_{A,c}(k) \geq t_{A,c})}{N_{ite}} & \text{if } t_{A,c} \geq 0 \\ 2 \cdot \frac{\sum_k I(T_{A,c}(k) \leq t_{A,c})}{N_{ite}} & \text{otherwise} \end{cases}$$

where the identity function $I(x)$ returns 1 if x is true and 0 otherwise.T he multiplication by 2 accounts for a two-tailed test, since we do not have a particular expectation regarding the pathway status (inhibited or activated).

3 COMBINING P_{NDE} AND P_{PERT} AND INTO A GLOBAL PATHWAY SIGNIFICANCE MEASURE.

After computing a p-value for both types of evidence, P_{NDE} and P_{PERT}, we need to combine these two probabilities into one global probability value, P_G, that will be used to rank the pathways and test the research hypothesis, that the pathway is significantly impacted in the condition studied. The probability that a pair of p-values, (P_{NDE}, P_{PERT}), is observed when the null hypothesis is true, can be computed based on the fact that, under the null hypothesis, a p-value is a uniformly distributed random variable on the interval $(0, 1)$. The surface of all theoretically possible values that the variables P_{NDE} and P_{PERT} can take is a square with unity area. The two probability values obtained for a given pathway P_i can be represented as a point within this square ($P_{NDE}(i)$, $P_{PERT}(i)$), as shown in Fig. 1.

$P_{PERT}(i)$). Since under the null hypothesis $P_{NDE}(i)$ and $P_{PERT}(i)$ are independent probabilities, they can be multiplied to give the joint probability of obtaining the observed number of DE genes *and* the observed perturbation at the same time. The geometrical locus of the points with the same joint probability is the hyperbola $P_{NDE}(i) \cdot P_{PERT}(i) = c$. The probability to obtain a set of p-values as extreme or more extreme than ($P_{NDE}(i)$, $P_{PERT}(i)$), is the area under and to the left of this hyperbola. The sought global probability P_G is the probability to have such a combination with a product less than or equal to that observed. Hence, P_G can be quantified as the ratio of the area under the curve divided by the entire area of the square (which is 1):

$$P_G = \int_0^c 1 \cdot dx + \int_c^1 \frac{1}{x} \cdot dx = c + c \cdot \ln x|_c^1 = c - c \cdot \ln c \quad (9)$$

In the example shown in Fig. 1, $P_{NDE}(i) = 0.2$ and $P_{PERT}(i) = 0.4$ which yields $P_G(i) = 0.282$. Eq. 9 can be used to calculate the constant c for any desired significance threshold α. For instance, for the the customary $\alpha = 0.05$, the product of the two individual probabilities can be calculated as $c = 0.0087$, a value which has been independently obtained by others (Loughin, 2004).

Since several pathways are tested simultaneously, we also need to consider adjusting the nominal $P_G(i)$ values for multiple comparisons. For the convenience of the user, the package implementing SPIA provides both Bonferroni- and FDR-corrected p-values.

4 SUPPLEMENTARY TABLES 1-10

REFERENCES

Loughin, T. (2004) A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics and Data Analysis,* **47** (3), 467 – 485.

Table 1. GSEA results on the Colorectal cancer dataset. Enrichement in cancer group. Output from R GSEA V 1.0.

	NOM p-val	FDR q-val	FWER p-val	FDR (median)	glob.p.val
Parkinsons..5020	0.008048	0.34709	0.192	0	0.155
Wnt signal..4310	0.01359	0.38628	0.354	0	0.14
Complement..4610	0.01961	0.31692	0.404	0	0.087
MAPK signa..4010	0.02115	0.17632	0.554	0.11785	0.011
Gap juncti..4540	0.02745	0.19836	0.657	0.14463	0.006
Axon guida..4360	0.02994	0.20308	0.484	0	0.03
Basal cell..5217	0.03571	0.23785	0.479	0	0.044
Colorectal..5210	0.03868	0.18203	0.528	0.12153	0.017
mTOR signa..4150	0.05253	0.16795	0.571	0.10938	0.004
Focal adhe..4510	0.06114	0.28562	0.466	0	0.07
ECM-recept..4512	0.06759	0.19884	0.513	0.13158	0.025
Regulation..4810	0.07968	0.23706	0.732	0.17157	0.01
Renal cell..5211	0.1053	0.34187	0.903	0.28226	0.015
Type II di..4930	0.1076	0.33158	0.848	0.26055	0.029
Melanogene..4916	0.1804	0.3619	0.942	0.3114	0.017

Table 2. GSEA results on the Colorectal cancer dataset. Enrichement in normal group. Output from R GSEA V 1.0.

	NOM p-val	FDR q-val	FWER p-val	FDR (median)	glob.p.val
Huntington..5040	0.1094	1	0.768	1	0.535
Dentatorub..5050	0.2189	1	0.896	1	0.592
SNARE inte..4130	0.2633	1	0.936	1	0.541
PPAR signa..3320	0.3247	1	0.96	1	0.525
Olfactory ..4740	0.3648	1	0.985	1	0.56
GnRH signa..4912	0.3908	0.93841	0.985	0.92687	0.446
Cell cycle..4110	0.4065	1	0.981	1	0.603
ErbB signa..4012	0.4345	0.83531	0.986	0.82639	0.351
Insulin si..4910	0.519	0.94025	0.996	0.96997	0.466
Thyroid ca..5216	0.5369	0.79603	0.997	0.81818	0.249
Ubiquitin ..4120	0.574	0.76453	0.999	0.77493	0.165
Tight junc..4530	0.5866	0.85592	0.996	0.87931	0.342
Phosphatid..4070	0.6321	0.73111	0.999	0.7517	0.091
Maturity o..4950	0.6604	0.72686	1	0.74598	0.058
Adipocytok..4920	0.6654	0.8032	0.999	0.82018	0.233

Table 3. SPIA results on the Vessels dataset

KEGG Pathway	P_{NDE}	P_{PERT}	P_G	$P_{G,FDR}$	$P_{G,FWER}$	Status
Antigen pr..4612	0.0067	0.0004	0.0000	0.0016	0.0016	Activated
Axon guida..4360	0.0002	0.0908	0.0002	0.0045	0.0090	Inhibited
Neuroactiv..4080	0.0006	0.1992	0.0012	0.0170	0.0514	Inhibited
Focal adhe..4510	0.0003	0.5364	0.0016	0.0170	0.0681	Inhibited
Wnt signal..4310	0.0008	0.4244	0.0032	0.0251	0.1356	Activated
Regulation..4810	0.0042	0.0948	0.0035	0.0251	0.1508	Activated
Type I dia..4940	0.0011	1.0000	0.0083	0.0469	0.3556	Inhibited
Complement..4610	0.0023	0.4812	0.0087	0.0469	0.3750	Activated
Notch sign..4330	0.0392	0.0468	0.0134	0.0579	0.5756	Activated
ECM-recept..4512	0.0024	0.7560	0.0135	0.0579	0.5789	Inhibited
Cytokine-c..4060	0.0453	0.2172	0.0553	0.2161	1.0000	Inhibited
Gap juncti..4540	0.0970	0.1236	0.0650	0.2331	1.0000	Inhibited
TGF-beta s..4350	0.0262	0.5224	0.0724	0.2396	1.0000	Inhibited
Tight junc..4530	0.2171	0.0700	0.0788	0.2421	1.0000	Inhibited
Adherens j..4520	0.0598	0.3112	0.0927	0.2659	1.0000	Activated

Table 4. ORA results on the Vessels dataset

KEGG Pathway	P_{NDE}	$P_{NDE,FDR}$	$P_{NDE,FWER}$
Axon guida..4360	0.0002	0.0065	0.0083
Focal adhe..4510	0.0003	0.0065	0.0131
Neuroactiv..4080	0.0006	0.0086	0.0257
Wnt signal..4310	0.0008	0.0089	0.0357
Type I dia..4940	0.0011	0.0091	0.0453
Complement..4610	0.0023	0.0150	0.1000
ECM-recept..4512	0.0024	0.0150	0.1050
Regulation..4810	0.0042	0.0225	0.1801
Antigen pr..4612	0.0067	0.0321	0.2886
TGF-beta s..4350	0.0262	0.1127	1.0000
Notch sign..4330	0.0392	0.1390	1.0000
Renal cell..5211	0.0405	0.1390	1.0000
MAPK signa..4010	0.0442	0.1390	1.0000
Cytokine-c..4060	0.0453	0.1390	1.0000
GnRH signa..4912	0.0502	0.1438	1.0000

Table 5. GSEA results on the Vessels dataset, enrichement in UA group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR(median)	glob.p.val
Renal cell..5211	0.002953	1	0.5845	0.90625	0.4725
Cell cycle..4110	0.02559	0.71623	0.638	0.53704	0.274
Huntington..5040	0.07707	0.80367	0.787	0.66667	0.334
Thyroid ca..5216	0.1374	0.69938	0.8915	0.64444	0.2555
SNARE inte..4130	0.1621	0.607	0.7885	0.51786	0.217
Gap juncti..4540	0.2102	0.9406	0.941	0.90344	0.432
Axon guida..4360	0.2205	0.75634	0.958	0.74571	0.285
Maturity o..4950	0.224	0.80301	0.8865	0.74839	0.3435
Melanogene..4916	0.2933	0.80577	0.9755	0.79091	0.337
Focal adhe..4510	0.344	0.83775	0.958	0.82857	0.3575
Long-term ..4730	0.3685	0.78456	0.9805	0.77679	0.307
Tight junc..4530	0.3835	0.91601	0.9555	0.88176	0.4235
TGF-beta s..4350	0.4477	0.78065	0.9815	0.78733	0.297
ECM-recept..4512	0.4477	0.90872	0.9905	0.91143	0.418
PPAR signa..3320	0.5045	0.76883	0.994	0.78628	0.256

Table 6. GSEA results on the Vessels dataset, enrichement in UV group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR(median)	glob.p.val
Insulin si..4910	0.01515	0.47717	0.7915	0.39286	0.1165
Type II di..4930	0.06825	0.74158	0.7255	0.61111	0.303
Toll-like ..4620	0.07475	1	0.6295	0.81481	0.445
Parkinsons..5020	0.1381	0.60681	0.779	0.52381	0.202
Neuroactiv..4080	0.1388	0.42792	0.8725	0.39286	0.039
ErbB signa..4012	0.1756	0.66679	0.9595	0.64706	0.218
Type I dia..4940	0.2022	0.39259	0.8285	0.36667	0.04
Antigen pr..4612	0.2377	0.46468	0.8265	0.44	0.091
MAPK signa..4010	0.2721	0.74577	0.9815	0.73333	0.2695
Adipocytok..4920	0.2964	0.69667	0.9845	0.69565	0.2325
Natural ki..4650	0.3005	0.67092	0.9695	0.65812	0.2045
Cytokine-c..4060	0.338	0.69065	0.986	0.6875	0.1915
Alzheimers..5010	0.5056	0.89088	0.993	0.91124	0.401
Taste tran..4742	0.5474	0.74295	0.993	0.76389	0.155
Epithelial..5120	0.562	0.76919	0.993	0.78571	0.2225

Table 7. SPIA results on the LaborM dataset

KEGG Pathway	P_{NDE}	P_{PERT}	P_G	$P_{G,FDR}$	$P_{G,FWER}$	Status
Cytokine-c..4060	0.0000	0.0000	0.0000	0.0000	0.0000	Activated
ErbB signa..4012	0.0000	0.0112	0.0000	0.0001	0.0002	Activated
Jak-STAT s..4630	0.0000	0.2140	0.0000	0.0004	0.0011	Activated
Epithelial..5120	0.0044	0.0024	0.0001	0.0015	0.0059	Activated
Complement..4610	0.0003	0.0740	0.0003	0.0023	0.0113	Inhibited
MAPK signa..4010	0.0011	0.4076	0.0038	0.0288	0.1726	Activated
Toll-like ..4620	0.0007	0.9344	0.0058	0.0370	0.2591	Activated
Adipocytok..4920	0.0063	0.4524	0.0195	0.1099	0.8793	Activated
PPAR signa..3320	0.0044	1.0000	0.0284	0.1218	1.0000	Inhibited
TGF-beta s..4350	0.0408	0.1084	0.0284	0.1218	1.0000	Activated
Insulin si..4910	0.0159	0.2944	0.0298	0.1218	1.0000	Inhibited
Type II di..4930	0.0857	0.1668	0.0750	0.2813	1.0000	Inhibited
Thyroid ca..5216	0.1189	0.1528	0.0910	0.2897	1.0000	Inhibited
Wnt signal..4310	0.1143	0.1828	0.1017	0.2897	1.0000	Inhibited
Circadian ..4710	0.0774	0.2704	0.1019	0.2897	1.0000	Inhibited

Table 8. ORA results on the LaborM dataset

KEGG Pathway	P_{NDE}	$P_{NDE,FDR}$	$P_{NDE,FWER}$
Cytokine-c..4060	0.0000	0.0000	0.0000
Jak-STAT s..4630	0.0000	0.0002	0.0004
ErbB signa..4012	0.0000	0.0005	0.0014
Complement..4610	0.0003	0.0032	0.0130
Toll-like ..4620	0.0007	0.0067	0.0335
MAPK signa..4010	0.0011	0.0081	0.0485
PPAR signa..3320	0.0044	0.0249	0.1989
Epithelial..5120	0.0044	0.0249	0.1989
Adipocytok..4920	0.0063	0.0315	0.2833
Insulin si..4910	0.0159	0.0715	0.7150
Natural ki..4650	0.0283	0.1158	1.0000
TGF-beta s..4350	0.0408	0.1531	1.0000
Renal cell..5211	0.0521	0.1714	1.0000
ECM-recept..4512	0.0533	0.1714	1.0000
Circadian ..4710	0.0774	0.2263	1.0000

Table 9. GSEA results on the LaborM dataset, enrichement in TL group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR(median)	glob.p.val
Epithelial..5120	0.00497	0.18574	0.1105	0	0.0965
MAPK signa..4010	0.005066	0.13574	0.3995	0	0.007
Cytokine-c..4060	0.008214	0.22727	0.318	0	0.0585
ErbB signa..4012	0.01091	0.16509	0.3675	0	0.0185
TGF-beta s..4350	0.01094	0.29343	0.2815	0	0.1215
Jak-STAT s..4630	0.01132	0.14696	0.3835	0	0.013
VEGF signa..4370	0.02198	0.18533	0.3405	0	0.0305
Adipocytok..4920	0.02402	0.14489	0.491	0	0.006
Complement..4610	0.04893	0.1504	0.4675	0	0.0085
Toll-like ..4620	0.06592	0.16007	0.5515	0.11176	0.006
Fc epsilon..4664	0.07407	0.16072	0.5785	0.11144	0.0035
Insulin si..4910	0.08918	0.2541	0.828	0.20956	0.004
Type II di..4930	0.09164	0.25983	0.77	0.20276	0.0105
Renal cell..5211	0.09323	0.26766	0.7935	0.20879	0.0105
Natural ki..4650	0.1095	0.16695	0.613	0.11728	0.0035

Table 10. GSEA results on the LaborM dataset, enrichement in TNL group. Output from R GSEA V 1.0.

KEGG Pathway	NOM p-val	FDR q-val	FWER p-val	FDR(median)	glob.p.val
Parkinsons..5020	0.2297	1	0.88	1	0.7075
Melanogene..4916	0.268	1	0.985	1	0.8135
Basal cell..5217	0.3283	1	0.9955	1	0.566
Amyotrophi..5030	0.4055	1	0.9955	1	0.755
Wnt signal..4310	0.4554	1	0.9995	1	0.628
Phosphatid..4070	0.4956	1	0.9995	1	0.502
Long-term ..4730	0.5726	0.97732	1	0.96571	0.46
Thyroid ca..5216	0.6073	0.94022	1	0.93914	0.423
Tight junc..4530	0.6633	0.90019	1	0.90794	0.3625
Gap juncti..4540	0.8018	0.93991	1	0.9602	0.4185
Olfactory ..4740	0.9208	1	1	1	0.9045
Regulation..4140	0.9756	1	1	1	0.964
Taste tran..4742	0.9902	0.98757	1	1	0.759