*Published online 27 November 2006* *Nucleic Acids Research, 2006, Vol. 34, No. 22* ***e151***
*doi:10.1093/nar/gkl766*

# Absolute enrichment: gene set enrichment analysis for homeostatic systems

**Vishal Saxena*, Dennis Orgill[1] and Isaac Kohane[2]**

Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [1]Surgery, Harvard Medical School, Boston, Chief, Burns unit, Brigham and Women's Hosptial, Boston, MA, USA and [2]Pediatrics and Health Sciences and Technology, Harvard Medical School, Boston, Director, Children's Hospital Informatics Program, Boston, MA, USA

Received April 18, 2006; Revised July 25, 2006; Accepted September 28, 2006

## ABSTRACT

**The Gene Set Enrichment Analysis (GSEA) identifies sets of genes that are differentially regulated in one direction. Many homeostatic systems will include one limb that is upregulated in response to a downregulation of another limb and vice versa. Such patterns are poorly captured by the standard formulation of GSEA. We describe a technique to identify groups of genes (which sometimes can be pathways) that include both up- and down-regulated components. This approach lends insights into the feedback mechanisms that may operate, especially when integrated with protein interaction databases.**

## INTRODUCTION

The Gene Set Enrichment Analysis (GSEA) (1) is a powerful technique for elucidating various groups of genes that may be important from gene expression data (2,3). However, one drawback of the current implementation of GSEA is that it only identifies gene sets regulated in one direction. This is problematic for several kinds of physiological processes. For example, in most homeostatic processes, when one component of the process is upregulated, there is a controlling downregulation in response and conversely downregulation of one component leads to an upregulation of another component all in order to maintain constancy of a particular set point. Figure 1 illustrates this point.

We introduce a simple and novel yet very powerful methodology that 'looks' at combined up- and down-regulated expression. This technique allows the computational highlighting of groups of genes or systems that may not be as easily identified through the GSEA algorithm. This methodology, called the Absolute Enrichment (AE) was applied to a dataset obtained from GEO Datasets (http://www.ncbi.nlm.nih.gov/geo/), (4,5). This dataset was obtained from patients who underwent hysterectomy or abdominal myomectomy for symptomatic uterine fibroids (6).

The GSEA algorithm is easily generalized to any procedure that results in a ranking of genes in an expression experiment. Although Mootha (1) implemented a 'signal-to-noise ratio' for his first implementation, other statistics more relevant to particular studies can be used in implementing the GSEA. For example, in comparisons of two groups that are a matched time series, the paired *t*-test can be the more appropriate test, and the paired *t*-statistic is used to rank the genes in such systems. The use of the paired *t*-statistic (or any other statistic) is driven by the nature of the experiments and is merely the ordering metric and does not address the challenge of identifying bidirectionally perturbed groups of genes. The method of AE, as described below, does not however depend on any particular metric (for the same reasons that GSEA is metric independent).

Gene set enrichment is most often used when one-gene-at-a-time analysis reveals no significant differential expression but when a set of genes might. In this paper, we describe the Absolute Enrichment (analysis) or AE which takes both up and down regulations into account (using the absolute signal-to-noise ratio or absolute SNR as the ranking metric) as well as the standard GSEA for comparison.

## MATERIALS AND METHODS

We will first describe Mootha's (1) original approach to the implementation of the GSEA. Next, we will describe our AE approach using the absolute SNR as the ranking metric.

The analyses presented in this paper were conducted on datasets that were generated from RNA hybridizations to HG-U133A Affymetrix human genechips. Thus, we created gene sets for this genechip.

### Standard GSEA using the signal-to-noise ratio SNR as the ranking metric

Given an expression dataset that has an 'affected' group and a 'control' group, one starts by taking (for each gene or probeset) the average of all the controls and the average for all the affecteds. The difference between the average value

*To whom correspondence should be addressed. Tel: +1 6172302904; Fax: +1 6177326387; Email: vishal@mit.edu

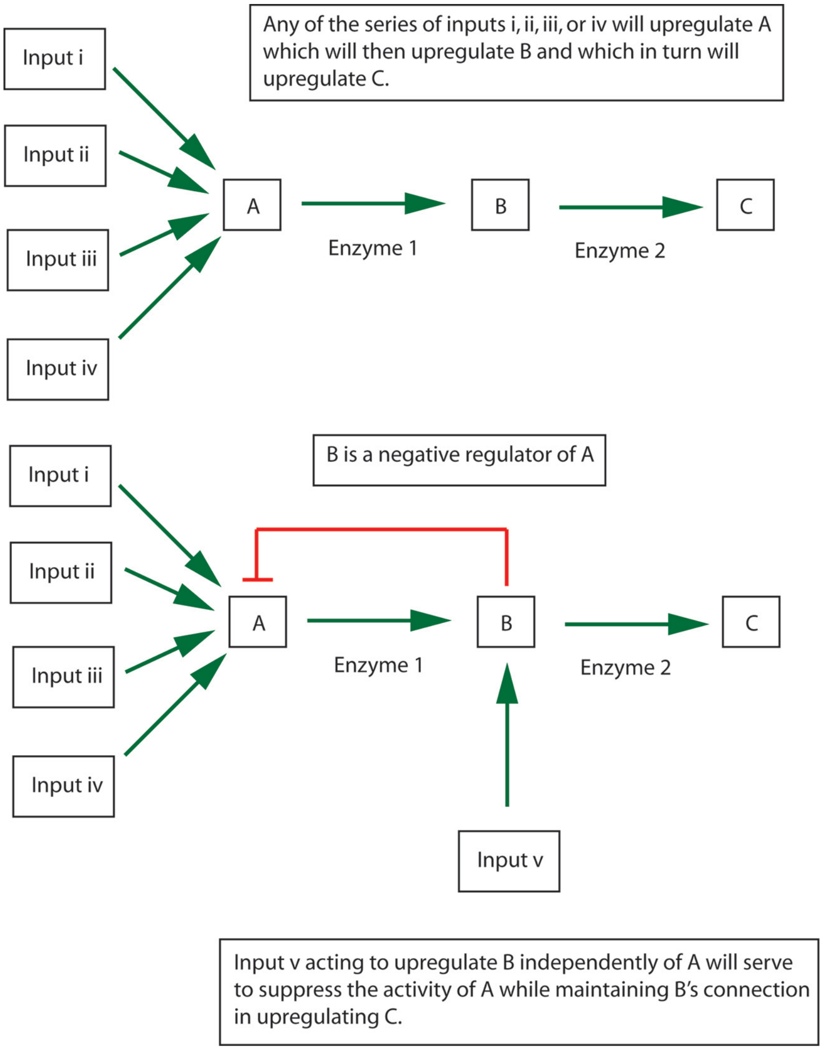
Input i
Input ii
Input iii
Input iv
Any of the series of inputs i, ii, iii, or iv will upregulate A which will then upregulate B and which in turn will upregulate C.
A
B
C
Enzyme 1
Enzyme 2
Input i
Input ii
Input iii
Input iv
B is a negative regulator of A
A
B
C
Enzyme 1
Enzyme 2
Input v
Input v acting to upregulate B independently of A will serve to suppress the activity of A while maintaining B's connection in upregulating C.

**Figure 1.** Pathways are in essence made up of proteins and other 'elements' that affect downstream elements (which may or may not show feedback on upstream elements). Most pathways have some feedback mechanism from downstream to upstream elements. When the input doesn't come from the (conceptualized) start of the pathway (there is really no start point—the pathway is what it is), upstream elements from the input may be downregulated through negative feedback as depicted here.

of the affected and the average value of the control group is taken and then divided by the sum of the standard deviations of each group to give us the signal-to-noise ratio statistic. The largest positive number then is the most upregulated. GSEA as described by Mootha (1) only looks at perturbation in one direction (e.g. upregulation) and the genes are ordered by signal-to-noise ratio (giving us an ordering based on upregulation). Each row contains the data pertaining to a single probeset or gene and thus each row is reordered based on the ranking metric (SNR in Mootha's algorithm).

Next the Kolmogorov Smirnov statistic is calculated according to the following equations:

$$XN = -\sqrt{\frac{G}{N-G}}$$

if the gene is not part of the gene set, and

$$X = \sqrt{\frac{N-G}{G}}$$

if the gene is part of the gene set, where $G$ is the number of genes in the gene set and $N$ is the total number of genes in the dataset.

Next a running sum is obtained on the (reordered) dataset. The maximum of the running sum is the value of the enrichment score.

Essentially the enrichment score is a measure of the 'enrichment' of the gene set at the top of the list of (the reordered) genes. The highest enriched gene set then is the most significantly differentially expressed gene set in the system under study.

This gene set is then tested for significance. If the columns (pertaining to sample conditions belonging to either control or affected) are randomly shuffled over the course of 1000 permutations, the number of times each gene set comes to the top divided by 1000 gives us an estimate of the $P$-value (1). If the $P$-value is higher than a predefined value (e.g. >0.05) then that gene set is not significantly differentially expressed in the analysis.

### The ranking metric in the GSEA and other metrics (e.g. the *t*-statistic)

The ranking metric used in the GSEA is not fixed to a certain type. Mootha *et al*. (1) averaged each data class or condition and took the difference (and then divided by the sum of the standard deviations of each data class) to obtain the SNR. Mootha's (1) data are replicates at the same time point. There are no pairwise comparisons to be made. In the case of a time series where we have a paired control and affected sample at each time point, the paired *t*-statistic may be more appropriate. The AE just like the GSEA is independent of ranking metric—either the SNR or the paired *t*-statistic or essentially any metric that discriminates between the 'affected' and 'control' classes can be used.

To be able to generate enough permutations to run the *P*-test (for significance) without generating repeats, we require a relatively large number of affected and control samples. Eight samples (or more) in each class are usually enough to meet this criteria.

### AE

Rather than ordering the genes by SNR, a value that goes from negative to positive, we simply take the absolute value thereby lumping together up- and down-regulated components of the homeostatic system. The datasets analyzed in this paper used this absolute value statistic in implementing the AE.

## RESULTS

### Applications of the AE

Datasets available at GEO datasets (4,5) at the NCBI website were analyzed using our AE analysis. One particularly illuminating dataset (using the HG-U133A genechip) measured gene expression in uterine fibroids (6). On running the AE analysis on this dataset, we found that the gene set c7_U133_probes from Mootha's original gene set grouping was topmost ranked (see Figure 2 for the expression pattern of this gene set). This gene set, however, was not the top ranked gene set in either the up-regulated analysis or the down-regulated analysis. The top ranked gene set in the standard GSEA analysis was the OXPHOS gene set.

As in Mootha's work (1), the SNR metric was used to rank the genes (however, absolute values were taken). The c7 gene set is number three in the down-regulated analysis, number five in the up-regulated (or standard GSEA) analysis, and then is top ranked in the AE analysis. The SNRs for the genes in this gene set are given in Figure 2.

The c7 gene set was then entered into EASE (a utility for calculating overrepresentation of Gene Ontology (GO) annotations in lists of genes) (7,8), (http://david.niaid.nih.gov/david/ease.htm) to see which GO categories are enriched. The top ranked 'biological process' GO category is 'cell adhesion', the top ranked 'molecular function' GO category is 'extracellular matrix structural constituent' and the top ranked 'cellular component' GO category is 'extracellular matrix'. Table 1 lists the top ranked GO categories pertaining to each of these GO categories separately (the aggregate rank obtained through EASE is given in the third column).

Nine other datasets were studied and four of these showed a different gene set appearing as the top ranked AE gene set compared with either the up-or down-regulated analyses. These results are presented in Table 2.

## DISCUSSION

### The AE may bring important insights to gene expression datasets

The AE can identify important gene sets that may not be identified through either the up- or down-regulated analyses. For example, in the uterine fibroid dataset, the c7 gene set was top ranked. This gene set was then analyzed through EASE. This gene set was found to be enriched for all three GO categories pertaining to the Extracellular matrix (Table 1 lists the rankings of the three GO categories derived from this gene set through the use of EASE).

According to (9) 'Uterine fibroids (correctly called leiomyomas or myomas) are benign myometrial neoplasms enriched in extracellular matrix (ECM)'. This paper lends support to the use of the AE, because it shows us that the gene set that has obtained top rank in the AE is made up of components that others have shown to be important in that system.

To further validate the c7 group of (possibly) coregulated genes, we studied the up- and down-regulated components

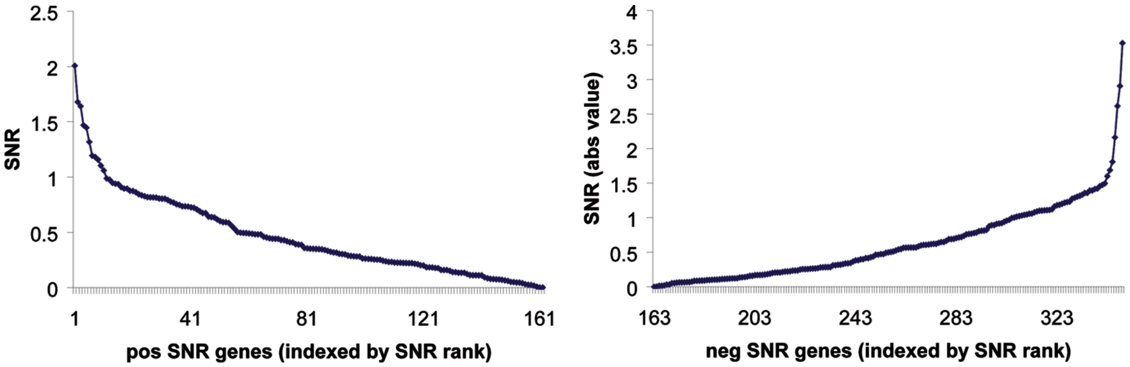

**Figure 2.** SNR scores from the uterine fibroid paper. The right graph is an absolute value plot (to show more clearly the symmetry in the way the gene set responded to enrichment).

of this gene set to see if the positive or negative expression values in the leiomyomas versus control have previously been shown to have similar signaling either in leiomyomas of the uterus or in other similar and analogous pathologies.

To perform the above analysis, all the genes from this gene set were run through EASE. The highest rated gene set genes were then sub-extracted (pertaining to ECM). These were then segregated into positive and negative ones (with SNR). COL4A5 was highest positive. It is reported in uterine leiomyomas by Catherino (10), and in esophageal leiomyomas by Aszodi (11). LAMB2 is another gene with altered expression in ECM diseases (11). It is the second most downregulated in the ECM set. The laminin, LAMB2, is downregulated in the progression from Prostatic Intraepithelial Neoplasia (PIN) to Prostate Cancer (PC) (12).

**Table 1.** Top 24 GO categories (separately ranked by the GO categories: Biological process, molecular function and cellular component) for the top ranked gene set in the AE analysis from the uterine fibroid dataset analyzed through EASE

| Rank | Top groups from GO category biological process | Overall rank within all three GO categories |
|---|---|---|
| 1 | Cell adhesion | 2 |
| 2 | Cell communication | 7 |
| 3 | Morphogenesis | 9 |
| 4 | Development | 12 |
| 5 | Organogenesis | 13 |
| 6 | Cell growth | 16 |
| 7 | Regulation of cell growth | 17 |
| 8 | Cellular process | 18 |
| 9 | Skeletal development | 19 |
| 10 | Regulation of cellular process | 21 |
| 11 | Regulation of biological process | 22 |
| Rank | Top groups from GO category molecular function | Overall rank within all 3 GO categories |
| 1 | Extracellular matrix structural constituent | 3 |
| 2 | Cell adhesion molecule activity | 6 |
| 3 | Structural molecule activity | 8 |
| 4 | Extracellular matrix structural constituent conferring tensile strength | 11 |
| 5 | Calcium ion binding | 15 |
| 6 | Glycosaminoglycan binding | 20 |
| 7 | Insulin-like growth factor binding | 23 |
| Rank | Top groups from GO category cellular component | Overall rank within all 3 GO categories |
| 1 | Extracellular matrix | 1 |
| 2 | Extracellular | 4 |
| 3 | Collagen | 5 |
| 4 | Basement membrane | 10 |
| 5 | Fibrillar collagen | 14 |
| 6 | Extracellular space | 24 |

Two genes that were upregulated in the leiomyoma were COL5A1 and LUM and these have previously been shown to be two of the 'the top discriminators' in osteosarcomas (13).

We next looked at the second most important EASE GO category (within the c7 gene set) and this is 'cell adhesion' (the highest ranked 'cellular process' category). In this category, TGF-β1 is upregulated and vinculin is downregulated. In cancer, an increase in TGF-β has been shown to lead to a decrease in vinculin (14), a cytoskeletal molecule. Although, leiomyomas are certainly not cancerous, they are tumors and therefore closer to cancer type tissue than normal tissue.

Cyr61 is a gene that is downregulated in the Cell adhesion GO category. This gene has been previously reported to have been downregulated in other microarray experiments pertaining to uterine leiomyomas (15). Another gene, CSPG2 was upregulated in our system. This gene has been seen to be upregulated in extraskeletal myxoid chondrosarcoma (16). The authors of this paper further state that 'CSPG2 is a protein that may play a role in intercellular signaling and in connecting cells with the extracellular matrix. Yoon *et al.* showed that CSPG2 is directly transactivated by p53′ (16). We then checked for the expression of p53 in our system and found that of the 29 probesets with p53 in their title, 19 were upregulated.

We next studied the whole c7 gene set for up and downregulation vis-à-vis other work in the literature. The most down-regulated gene in our system in the c7 gene set is ANXA1. This gene has been shown in other microarray experiments to be downregulated (17). Further the authors in (17) mention that XLKD1 was downregulated in their study and we see that in the dataset we looked at XLKD1 was also downregulated.

Vanharanta *et al.* (18) used cDNA microarrays to study uterine fibroids. They showed that LTBP1 was downregulated as it is in our system in the c7 gene set.

The above analysis lends further support for the validity of the AE.

**Table 2.** AE, up and down-regulated analyses conducted on 10 datasets

| Dataset analysis | AE top gene set | Up-regulated top gene set | Down-regulated top gene set | Up[a] | Down[b] |
|---|---|---|---|---|---|
| Uterine fibroid | c7_U133_probes | OXPHOS | c9_U133_probes | 5 | 3 |
| COPD | Human_mitoDB | MAP00480_Glutathione_metab | Human_mitoDB | 116 | 1 |
| Endothelin fibroblasts | c7_U133_probes | c7_U133_probes | MAP00100_Sterol_biosyn | 1 | 17 |
| Male female hypothal | c7_U133_probes | c34_U133_probes | c9_U133_probes | 194 | 2 |
| Lung cancer motexafin | OXPHOS | Mitochondr | c11_U133_probes | 3 | 6 |
| Sarcoma and hypoxia | c24_U133_probes | c7_U133_probes | c24_U133_probes | 169 | 1 |
| Squamous lung cancer | Human_mitoDB | Human_mitoDB | c9_U133_probes | 1 | 187 |
| Tumor cell topoisom. | Human_mitoDB | OXPHOS | c4_U133_probes | 2 | 34 |
| Breast cancer | OXPHOS | c2_U133_probes | c9_U133_probes | 8 | 3 |
| Bladder SMC stretch | c7_U133_probes | OXPHOS | c7_U133_probes | 11 | 1 |

[a] Refers to the ranking of the top AE gene set on the up-regulated list.
[b] Refers to the ranking of the top AE gene set on the down-regulated list.

### Caveat in conclusions drawn through the use of EASE about GO categories

GO only annotates ~60% of all genes for *Homo sapiens* (19). This creates a bias in any analysis that uses GO. This bias may be present in the validation step that we conducted through the use of EASE since EASE makes use of GO derived categories.

### Down-regulated gene sets are just as important to study as up-regulated and absolute enrichment gene sets

It should be noted that the traditional GSEA talks only of up-regulated gene sets. While upregulated groups of (possibly coregulated) genes are important to study, it may be equally important to look at down-regulated gene sets. When gene sets are connectable to GO biological process categories (those that come closest to defining pathways), those gene sets that are down-regulated, then talk of (possible) pathways that may be turned off in the system under study.

It would be important to know which groups of genes are turned off just as it is important to know those that are turned on. For example, the turning off of tumor suppressor genes has profound effects on the cell phenotype as it leads to the development of cancer, and this fact is probably just as important as the turning on of oncogenes in the cell.

We can think of an example outside biology that can also illustrate the importance of looking at downregulation. When the brakes of an automobile are working (and therefore on), the automobile can be safely driven. When the brakes are not working (and therefore off and not able to function), the automobile is highly dangerous to drive and is prone to accidents. Taking this analogy further to see how homeostatic mechanisms can be captured, we see that if the automobile is running at very high speeds it could be that this is a normal condition (fast driving). However, when we combine this with absent brakes, we obtain the homeostatic picture where the automobile is running fast because another component that controls or suppresses its function is now defective and 'downregulated'.

### Gene sets are not necessarily pathways and GO gene categories should not be mixed

At the GO website, it is stated that 'A biological process is a recognized series of events or molecular functions. A biological process is not equivalent to a pathway, although some GO terms do describe pathways.' Thus, although some GO biological process categories may be deemed as pathways, most GO categories are much simpler. Thus, we should not necessarily draw pathway conclusions from the use of GO categories (although some may pertain to pathways).

### Absolute enrichment is not asymmetric because of the concurrent use of down regulations along with up regulation

It may seem that up- and down-regulated data are asymmetric. It may seem that there is no theoretical limit to up regulation, while there may seem to be a limit to down regulation since we cannot go below zero in gene expression. However, this is not necessarily the case. Up and down regulations are defined relative to another condition. If we take up regulation to be the affected minus the control while taking down regulation to be control minus affected, we see that there does not need to be a limit to the level of expression in the control. Thus, the definition of up and down regulated is arbitrary depending on whether we subtract control from the affected or vice versa.

We further checked the uterine fibroid dataset for any asymmetry in the control versus the affected along numerous metrics such as largest absolute expression in the control and affected, largest absolute difference value between affected and control and vice versa, and largest signal-to-noise ratio between affected and control and vice versa. Along each of these metrics, we found consistent symmetry between the down-regulated expression values and up-regulated expression values. Thus, the AE statistic does not suffer from any type of asymmetry because of the use of the down-regulated limb in our analysis.

The AE statistic is usually not more likely to extract gene sets from the middle of the ranked list (by enrichment score) of gene sets in either the up-regulated or the down-regulated rankings

We also checked to see if the AE statistic is likely to pick up those gene sets that are not near the top of either the up- or down-regulated gene set rankings. The results of our analyses on nine additional datasets (and also the results from the uterine fibroid dataset) are shown in Table 2. We note that the AE statistic often captures gene sets that are typically near the top or bottom of either the up- or down-regulated gene set lists (204 gene sets were run) but sometimes far enough away from the ends to be missed.

We note that even if the top ranked AE gene sets are far from the up- or down-regulated lists, this is not necessarily a sign of any error since it could just be that in the system under study most gene sets that were defined turned out to be heavily either up or down regulated and those that were both up and down regulated therefore found themselves in the center.

Further, we also note that the AE is more likely to capture higher levels of differential expression precisely because it makes up- and down-regulated components of the gene expression symmetric thus capturing more extremely differentially expressed genes (depending on the gene sets) vis-à-vis just up- or down-regulated analyses which may be more likely to capture genes from the middle of the reordered lists of genes (reordered by the ranking metric such as the absolute SNR).

Although we found a four out of nine or 44% chance of discovering new gene sets in our AE analysis that were not top ranked in either the up- or down-regulated analyses, we feel that this number itself may be reflective of how gene sets themselves have traditionally been defined. Many of the 'already-made' gene sets that we obtained from the literature and the ones that we constructed ourselves were created from clustering techniques applied to gene expression datasets that made use of solely either up- or down-regulated gene groups. Thus, if the gene sets themselves are created from such solely up- or solely down-regulated clusters, then they may tend to only capture those aspects of other datasets to which they are applied in analyses. We suspect that if gene sets are created using bidirectional clustering techniques then we expect the AE to capture these gene sets more often.

This line of thought also creates a justification for the use of bidirectional clustering in capturing homeostatic systems in analyzing gene expression datasets.

## ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by Children's Hospital Informatics Program. Dr. Kohane is partially supported by National Institutes of Health National Center for Biomedical Computing grant 5U54LM008748-02.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Mootha,V.K. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.*, **34**, 267–273.
2. Zhou,Y. *et al.* (2005) *In silico* gene function prediction using ontology-based pattern identification. *Bioinformatics*, **21**, 1237–1245.
3. Kim,S.Y. and Volsky,D.J. (2005) PAGE: Parametric Analysis of Gene set Enrichment. *BMC Bioinformatics*, **6**, 144.
4. Barrett,T. *et al.* (2005) NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res.*, **33**, D562–D566.
5. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
6. Hoffman,P.J. *et al.* (2004) Molecular characterization of uterine fibroids and its implication for underlying mechanisms of pathogenesis. *Fertil. Steril.*, **82**, 639–649.
7. Dennis,G., Jr *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, p3.
8. Hosack,D.A. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.
9. Walker,C.L. and Stewart,E.A. (2005) Uterine fibroids: the elephant in the room. *Science*, **308**, 1589–1592.
10. Catherino,W. *et al.* (2004) Gene expression studies in leiomyomata: new directions for research. *Semin. Reprod. Med.*, **22**, 83–90.
11. Aszodi,A. *et al.* (1998) Mouse models for extracellular matrix diseases. *J. Mol. Med.*, **76**, 238–252.
12. Ashida,S. *et al.* (2004) Molecular features of the transition from prostatic intraepithelial neoplasia (PIN) to prostate cancer: genome-wide gene-expression profiles of prostate cancers and PINs. *Cancer Res.*, **64**, 5963–5972.
13. Baird,K. *et al.* (2005) Gene expression profiling of human sarcomas: insights into sarcoma biology. *Cancer Res.*, **65**, 9226–9235.
14. Akhurst,R.J. and Derynck,R. (2001) TGF-beta signaling in cancer--a double-edged sword. *Trends Cell. Biol.*, **11**, S44–S51.
15. Lee,E.J. *et al.* (2005) Profiling of differentially expressed genes in human uterine leiomyomas. *Int. J. Gynecol. Cancer*, **15**, 146–154.
16. Subramanian,S. *et al.* (2005) The gene expression profile of extraskeletal myxoid chondrosarcoma. *J. Pathol.*, **206**, 433–444.
17. Ahn,W.S. *et al.* (2003) Targeted cellular process profiling approach for uterine leiomyoma using cDNA microarray, proteomics and gene ontology analysis. *Int. J. Exp. Pathol.*, **84**, 267–279.
18. Vanharanta,S. *et al.* (2005) 7q deletion mapping and expression profiling in uterine fibroids. *Oncogene*, **24**, 6545–6554.
19. Lomax,J. (2005) Get ready to GO! A biologist's guide to the Gene Ontology. *Brief Bioinformatics*, **6**, 298–304.