BIOINFORMATICS

Vol. 23 ECCB 2006, pages e84–e90
doi:10.1093/bioinformatics/btl315

Discovering tightly regulated and differentially expressed gene sets in whole genome expression data

Chun Ye[1,*] and Eleazar Eskin[2]

[1]Bioinformatics Program, University of California, San Diego, La Jolla, CA 92093-0404, USA and
[2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093-0404, USA

ABSTRACT

Motivation: Recently, a new type of expression data is being collected which aims to measure the effect of genetic variation on gene expression in pathways. In these datasets, expression profiles are constructed for multiple strains of the same model organism under the same condition. The goal of analyses of these data is to find differences in regulatory patterns due to genetic variation between strains, often without a phenotype of interest in mind. We present a new method based on notions of tight regulation and differential expression to look for sets of genes which appear to be significantly affected by genetic variation.

Results: When we use categorical phenotype information, as in the Alzheimer's and diabetes datasets, our method finds many of the same gene sets as gene set enrichment analysis. In addition, our notion of correlated gene sets allows us to focus our efforts on biological processes subjected to tight regulation. In murine hematopoietic stem cells, we are able to discover significant gene sets independent of a phenotype of interest. Some of these gene sets are associated with several blood-related phenotypes.

Availability: The programs are available by request from the authors.

Contact: cye@bioinf.ucsd.edu

1 INTRODUCTION

1.1 Background

When microarrays were first introduced 10 years ago, researchers were attracted to the high throughput nature of the technology that enabled them to examine the activity of all genes in a cell simultaneously. The original promise of using this technology to quickly generate biologically relevant and verifiable hypotheses has yet to be fully realized. Ironically, much of the initial work has focused on the reduction of whole genome expression data to the identification of individual genes that exhibit differential expression. The most common such studies compare the expression levels of individual genes between two spatial or temporal conditions; or between diseased (e.g. cancerous) and normal cells. Various statistical tests are used to assess the significance of differential expression by comparing individual genes to the rest of the genes. Using these single-gene approaches, researchers have identified many genes critical to the function of various biological processes including the yeast cell cycle (Spellman *et al*., 1998) and the onset of human breast cancer (Perou *et al*., 2000).

*To whom correspondence should be addressed.

Because most biological processes involve the complex interaction and regulation of multiple genes, identifying differentially expressed sets of genes has important advantages over identifying individual genes. Many genes may individually exhibit marginal differential expression but may have a significant combined effect on phenotypic outcome. The most common gene set method directly extends single-gene approaches by:

(1) Ordering the complete list of genes, L, according to their evidence for differential expression.

(2) Examining the occurrence of a predefined gene set S to determine whether it is overrepresented in the top portion of the list, B, relative to the complete list L.

(3) Computing a P-value usually based on the Fisher's exact test or its large-sample approximation χ^2 test.

Numerous software and web sites perform this sort of analysis, most often by using Gene Ontology as the source of gene sets. Examples include GENMAPP (Dahlquist *et al*., 2002), CHIPINFO (Zhong *et al*., 2003), GOMINER (Zeeberg *et al*., 2003), ONTOTOOLS (Draghici *et al*., 2003), FUNCASSOCIATE (Berriz *et al*., 2003) and EASE (Hosack *et al*., 2003).

This approach is reasonable but has at least three shortcomings. First, by considering only those genes that belong to a gene set, others not in the set and their relative positions in the gene list are disregarded. Second, significant genes are arbitrarily selected using a threshold leading to different results from using different thresholds. Third, and most importantly, once B and S have been picked, the actual measure of differential expression (or other scores) associated with each gene is disregarded and any additional information, including correlation between gene expression levels is ignored (Pavlidis *et al*., 2004).

Exploiting the correlation structure between genes is an alternate approach of utilizing microarray data. It has been shown in previous studies that correlated expression levels between genes are directly associated with functional relationships such as physical interactions and common regulatory mechanisms (Eisen *et al*., 1998). As such, expression level correlation has been used to identify new functional modules and gene sets.

A promising new technique, gene set enrichment analysis (GSEA) uses binary phenotype information (e.g. samples belonging to two classes) and a new statistic similar to the Kolmogorov–Smirnov statistic to evaluate microarray data at the level of gene sets. The method avoids many of the problems of single-gene approaches by determining whether members of a gene set S tend to occur toward the top (or bottom) of the entire gene list L, in which case the gene

set is correlated with the phenotypic class distinction. This approach eliminates some of the single-gene biases and detects differentially expressed gene sets whose individual members are not necessarily differentially expressed under the single-gene model. The steps are (1) construct a ranked list L of genes based on correlation between their expression and class distinction using any suitable metric (signal-to-noise ratio), (2) given an a priori defined set of genes S, calculate an enrichment score (ES) that reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L, (3) estimate the statistical significance of the ES score by using an empirical phenotype-based permutation test procedure that preserves the complex correlation structure of the gene expression data and (4) correct for multiple testing by normalizing the ES score for each gene set to account for the size and then calculating the FDR for each normalized enrichment score.

The major advantage of GSEA is the ability to identify differentially expressed gene sets while preserving the correlation structure of the gene expression data avoiding a potential source of false positives. Tian *et al*. (2005) presents a potential solution by performing separate statistical tests based on two different but nevertheless related null hypotheses. Although statistically sound, such an approach makes the computation of a P-value difficult.

1.2 'Tightly regulated' gene sets

Recently, a new type of expression data is being collected which aims to measure the effect of genetic variation on gene expression of pathways (Morley *et al*., 2004; Bystrykh *et al*., 2005). In these datasets, expression data are collected from multiple strains of the same model organism under the same condition. The goal of analyses of these data is to identify differences in regulatory patterns due to genetic variation between strains. Previous gene set based methods, including GSEA, identify sets of genes which are highly variable across the strains and correlated with a phenotype of interest. An example of such a pathway is shown in Figure 1a. However, these types of datasets motivate a new notion of differential expression, tightly regulated gene sets, where differentially expressed genes are highly correlated within the strains. In this phenomenon, a set of genes is expressed at a different level of activity in each strain and may represent a different level of activity of a pathway. Figure 1b shows an example of a tightly regulated pathway.

In this paper, we formalize the notions of tight regulation and differential expression to generate two null hypotheses. First, we present a new method for identifying tightly regulated gene sets based on a correlation statistic. Second, we introduce a more general method of identifying differential expression to specifically handle whole genome expression data over different strains of a model organism without requiring a phenotype. From Figure 1, we see that our two notions of significance for a gene set capture different information. A differentially expressed gene set has contrasting vertical bands. Figure 1a and b both have a high ratio of genes differentially expressed while Figure 1c has only a few genes differentially expressed. A tightly regulated gene set has a smooth horizontal gradient. This pattern is observed in Figure 1b but not in Figure 1a and c.

We apply our methods to three datasets, two of which have been previously analyzed with other gene set based methods allowing us to directly compare our results. The first dataset is a diabetes expression profiling study of a binary phenotype [17 individuals with normal glucose tolerance (NGT) versus 18 individuals with Type 2 Diabetes Mellitus (DM2)]. The second dataset is an Alzheimer's expression profiling study of a continuous phenotype where 31 individuals were quantified by their MiniMental State Examination (MMSE) scores. These individuals are also clinically divided into four categories: control, incipient, moderate and severe. Finally, we apply our methods to a dataset containing expression profiles of hematopoietic stem cells (HSCs) from 22 strains of recombinant inbred mice.

2 METHODS

2.1 Hypothesis testing framework

The overall objective of our approach is to identify gene sets that are differentially expressed and tightly regulated. This gives us two null hypotheses as follows:

(1) Hypothesis D_0: The genes in a gene set show the same pattern of differential expression compared with the rest of the genes.

(2) Hypothesis T_0: The genes in a gene set show the same amount of correlation as the rest of the genes.

Notice two important differences in our framework. First, neither hypothesis explicitly requires phenotype information although many different notions of differential expression can be applied, including those involving phenotypes. Second, empirically these two hypotheses show very little correlation (data not shown).

2.2 Identifying tightly regulated pathways

We use the Kendall Tau to look for tightly regulated gene sets. Unlike the Pearson and Spearman correlations, there is an intuitive, graphical interpretation of the Kendall Tau. Given two genes, we create two ranked lists of the strains based on the expression levels of each gene. In graph theoretic terms, we are creating a bipartite graph with the strains representing the two sets of vertices. Each strain from one ranked list is connected to the same strain in the other ranked list by an edge.

Formally, given two genes i and j each with n expression values, one from each strain, the Kendall Tau is defined as:

$$\tau(Y_i, Y_j) = \frac{1 - 2c}{m(m-1)/2},$$

where c is the number of crossings in our bipartite graph and m is the number of strains. The specific statistic we use to evaluate gene sets is the sum of squares of all pairwise Kendall coefficients for a given gene set,

$$k = \sum_{i=1}^{n} \sum_{j=1}^{n} (\tau(Y_i, Y_j))^2.$$

To assess the significance of this statistic for a given gene set, we construct 2000 random gene sets with the same cardinality as our original gene set and recompute k. A P-value is then computed from the permutations.

2.3 Identifying differentially expressed pathways

2.3.1 ANOVA and variance For datasets that include categorical phenotype information, we use a linear mixed effects ANOVA model to rank the genes based on their F statistic. We use the MAANOVA R package and specify the gene-specific model as:$r_{igr} = G + S_i + I_i + \epsilon_{ir}$where the indicies track array (i), gene (g) and measurement (r); G is the average intensity associated with a particular gene; S is the effects associated with different samples or conditions (i.e. normal versus type 2 diabetes patients); I is the individual random effect of multiple individuals from a

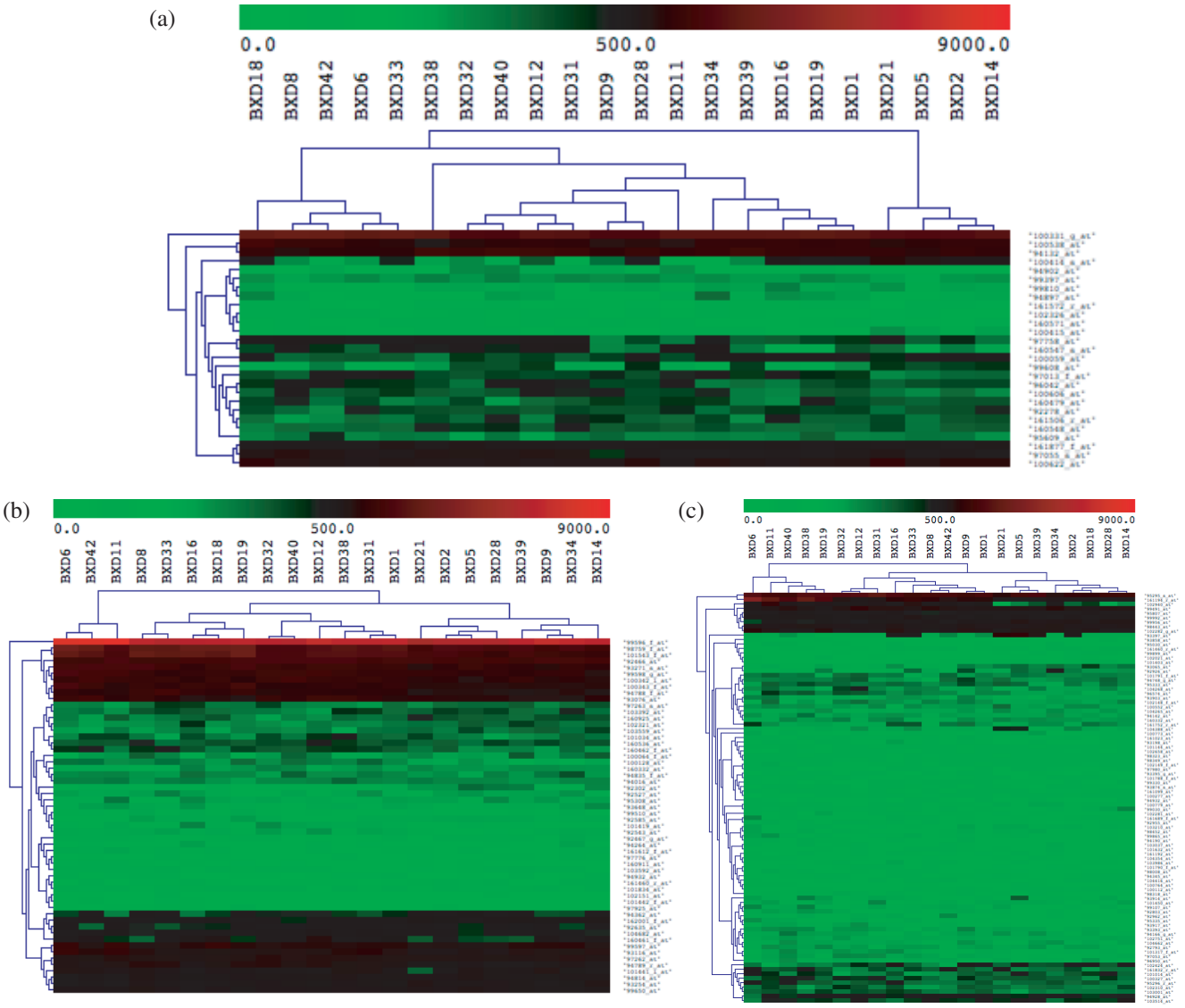

Fig. 1. The figure shows a two-way clustering of samples and genes in three different gene sets. (**a**) The GO oxygen and reactive oxygen species metabolism gene set is differentially expressed as evidenced by the sharp contrasting vertical lines but not tightly regulated as seen by the sharp contrasting horizontal lines. (**b**) The KEGG gap junction pathway is both differentially expressed ($P < 0.009$) and tightly regulated ($P < 0.0005$) in murine hematopoietic stem cells. (**c**) The KEGG cytokine–cytokine receptor interaction pathway is neither tightly regulated nor differentially expressed in murine hematopoietic stem cells.

particular sample (i.e. five diabetes patients); and $\in$ is the residual. We compute the following F statistic:

$$F = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{rss_1/df_1},$$

where rss_0, df_0 and rss_1, df_1 are the residual sums of squares and degrees of freedom for the null and alternative models, respectively. A P-value based on the normal distribution is returned by MAANOVA (Cui and Churchill, 2003).

For datasets that do not have clear categorical phenotype information, we look for differential expression based on the variance of the gene expression values of each gene. This is mostly applicable to whole genome expression data of different individuals from the same model organism.

2.3.2 The Mann-Whitney test To compute over-representation of a gene set in a ranked list, either derived from a linear ANOVA model or by computing the variance of the expression levels, we use the Mann–Whitney statistic. This non-parametric test uses the entire list to compute the statistic, an advantage over Fisher's exact test or the χ^2 approximation. Given a ranked list of genes, we expect genes from those gene sets exhibiting differential expression to cluster at the top of the list.

Formally, given a $m \times n$ matrix representing a microarray experiment with m strains and n genes, we first rank the genes based on their differential expression. We now take each gene set of interest S and compute the Mann–Whitney score where genes belonging to the set are compared with genes not in set. The alternate hypothesis is that those genes belonging to our gene set tend to show up higher on the list than other genes. The null hypothesis is that genes belonging to our gene set should be evenly distributed with respect

to all other genes. We use the R statistics package to compute the Mann–Whitney statistic and its associated *P*-values which are based on normal approximations (R Development Core Team, 2005).

3 RESULTS

To find correlated gene sets, we measure the amount of correlation between each pair of genes in the set. We use as a statistic the sum of the square of the correlation coefficients. For a gene set S, we first compute the correlation coefficients (Kendall Tau) for each pair of genes, c_{ij}. Then, we square each coefficient and sum the squares, $C = \sum_{i=1}^{n} \sum_{j=1}^{n} (c_{ij})^2$. Finally, we normalize this statistic by the number of elements in the gene set. To assess the significance we generate 2000 gene sets with the same number of genes as S, and recompute the statistic. A *P*-value is computed based on these iterations.

When we have groupings of individuals based on phenotype, for example, normal versus diabetic or severe versus incipient Alzheimer's, we use an ANOVA-based approach to look for differentially expressed gene sets. First, we construct a mixed-effect ANOVA model treating the effects due to disease state and variation among individuals as two fixed effects. Our null hypothesis is that gene expression does not vary between the groups. With this model, we use a F-statistic to measure how differentially expressed each gene is between the groups. If we do not have phenotypic groupings, as is the case for our murine expression data, we use the variance to measure differential expression for each gene. We rank each gene based on one of the two statistics for differential expression and use the Mann–Whitney statistic to detect significant gene sets, i.e. those with many genes occupying the top of the list.

Given these two statistical tests for tight regulation and differential expression, we can identify statistically significant gene sets using either statistic. When we computed these two statistics on the diabetes dataset, we were encouraged to see that there is very little correlation between the *P*-values ($\rho = 0.136$).

3.1 Example I: diabetes—NGT versus DM2

As a proof of concept, we first reanalyze the diabetes dataset originally reported in (Mootha *et al*., 2003). This dataset consists of whole genome expression profiles from 35 individuals, 17 of whom have NGT and 18 of whom have DM2. We consider the same 323 gene sets reported in (Mootha *et al*., 2003) compiled from various databases including BioCarta, GenMAPP and GO.

When we compare our methods with GSEA, we notice some interesting similarities and differences. Table 1 shows all significant gene sets using our method (sorted by the average rank of their *P*-values) and the GSEA method ($P < 0.05$). The gene set implicated in the original GSEA paper (Subramanian *et al*., 2005) and a follow up (Tian *et al*., 2005), OXPHOS, is the second highest scoring gene set using our approach while it is only marginally significant using the GSEA approach ($P < 0.01$). In addition to OXPHOS, we identify several more gene sets related to oxidative phosphorylation, including MAP00190 and KEGG: electron transport chain. Figure 2 shows a clustering of the top scoring gene sets. The represented biological processes include oxidative phosphorylation, proteasome degradation, carbon fixation and metabolism and cell–cell signaling. Some of these processes, including carbon fixation and pyruvate metabolism, are known to be differentially perturbed in diseased individuals (Randle *et al*., 1964). Others, especially the signaling pathways, are interesting potential biological targets because not only are they differentially expressed in diseased individuals, they are also tightly regulated, which suggests that a small number of regulatory elements govern the behavior of the pathway and the outcome of the phenotype.

Table 1. Significant diabetes gene sets

Gene Set	DE P	DE Rank	TR P	TR Rank	Avg Rank	GSEA P	GSEA Rank	Set Size
Electron_Transport_Chain	7.83E-13	1	5.00E-04	14	*7.5*	3.20E-02	17	88
VOXPHOS	3.45E-12	2	5.00E-04	14	*8*	1.20E-02	7.5	87
MAP00190_Oxidative_phosphorylation	2.80E-04	3	5.00E-04	14	*8.5*	3.10E-01	107	45
INSULIN_2F_UP	4.00E-04	4	5.00E-04	14	*9*	8.22E-01	267	214
GO_0005739	1.38E-02	10	5.00E-04	14	*12*	7.50E-01	242	170
chrebpPathway	1.38E-02	11	5.00E-04	14	*12.5*	4.00E-03	2.5	34
GLUCO	2.60E-02	13	5.00E-04	14	*13.5*	2.88E-01	99	32
MAP00500_Starch_and_sucrose_metabolism	2.66E-02	14	5.00E-04	14	*14*	1.82E-01	67	22
hsp27Pathway	2.91E-02	15	5.00E-04	14	*14.5*	2.64E-01	89	15
mta3Pathway	3.36E-02	16	5.00E-04	14	*15*	3.04E-01	105	15
Glycogen_Metabolism	3.77E-02	19	5.00E-04	14	*16.5*	2.26E-01	77	35
cell2cellPathway	9.72E-03	9	3.50E-03	38	*23.5*	6.00E-03	5.5	11
MAP00710_Carbon_fixation	1.36E-01	38	5.00E-04	14	*26*	6.16E-01	197	17
ccr3Pathway	1.53E-01	42	5.00E-04	14	*28*	3.20E-02	17	23
Proteasome_Degradation	9.83E-02	30	1.50E-03	31	*30.5*	9.26E-01	296	34
integrinPathway	1.30E-01	36	1.00E-03	28	*32*	2.00E-03	1	35
MAP00620_Pyruvate_metabolism	1.88E-01	50	5.00E-04	14	*32*	6.58E-01	211	31
MAP00020_Citrate_cycle_TCA_cycle	1.99E-01	54	5.00E-04	14	*34*	1.60E-01	62	18
MAP00330_Arginine_and_proline_metabolism	1.00E-01	31	4.00E-03	39.5	*35.3*	3.00E-02	15	39
RAP_DOWN	1.48E-01	41	1.50E-03	31	*36*	9.56E-01	311	227
GLUCOSE_DOWN	6.13E-02	23	1.10E-02	50	*36.5*	7.00E-02	25	155
Krebs-TCA_Cycle	2.25E-01	59	2.00E-03	34	*46.5*	6.18E-01	198	28
XINACT	9.13E-03	8	9.20E-02	90.5	*49.3*	2.86E-01	98	12
LEU_DOWN	3.10E-01	85	5.00E-04	14	*49.5*	3.46E-01	116	179

3.2 Alzheimer's data

Another dataset analyzed by Tian *et al*. (2005) considers a group of Alzheimer's patients. Blalock *et al*. (2004) performed gene expression experiments on 22 Alzheimer's disease (AD) patients and 9 controls. In addition to the clinically diagnosed categories of incipient, moderate and severe, each of these patients also have a Mini Mental Status Exam score which is a reliable index of AD-related cognitive status at a given point in time.

Using a database of gene sets compiled from BioCarta, KEGG, GENMAPP and GO (Tian *et al*., 2005), we perform two separate analyses to identify differentially expressed genes. First, we use ANOVA and the discrete phenotype to rank genes based on their differential expression with respect to the groupings. Second, we rank the genes based on the variance of their expression values. For each ranked list, we then use the Mann–Whitney test to identify top scoring gene sets. To identify tightly regulated gene sets, we use the same sum of squares correlation statistic as before and compute *P*-values based on 2000 permutations.

All methods used in this analysis detected an overwhelming number of significant gene sets, many of which appear to have housekeeping functions. Table 2 shows representative gene sets from the top 25 clusters and their respective statistics. We find some of the same gene sets identified in (Tian *et al*., 2005) including cation transporter activity (GO:0008324). Interestingly, an accompanying gene set, calcium ion transport (GO:0006816) is not discovered by our methods. Closer examination reveals that although the calcium ion transport gene set is slightly differentially expressed ($P < 0.096$), it is not tightly regulated ($P < 0.96$). In other words,

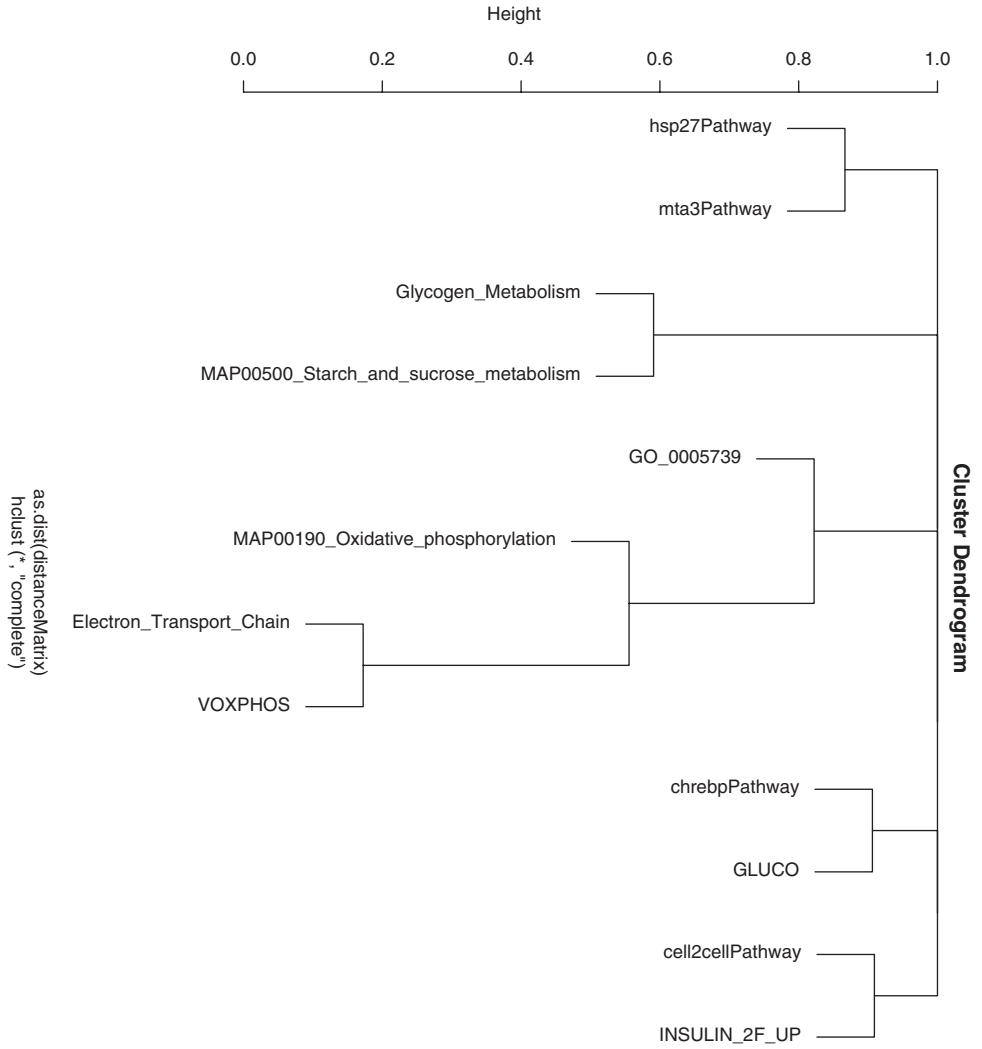

Fig. 2. This dendrogram of the top ranking gene sets. We see several clusters of gene sets roughly grouped by biological process, including oxidative phosphorylation, carbon fixation and metabolism, proteasome degradation, and cell–cell signaling.

although most of the proteins involved in calcium ion transport are channel proteins, the regulation might actually occur at the level of the ATP-dependent molecular transporter. This hypothesis is supported by the identification of numerous gene sets involved in metabolism including the cation-transporting ATPase activity gene set (GO:0019829). Several mitochondrial-related gene sets are also discovered supporting the pathogenesis of AD due to defects in these gene sets (Hinerfeld *et al*., 2004). Gene sets involving gluconeogenesis and glycolysis, two processes known to be involved in neurodegenerative disorders, are also among the top scoring sets.

Finally, of note is that many KEGG pathways involving neurodegenerative diseases, including the Alzheimer's pathway and the Neurodegenerative Diseases pathway are differentially expressed under one criterion (ANOVA) but not under another (variance). These pathways include many genes such as APP, APOE and PSEN whose activity has been well documented to be associated with Alzheimer's.

3.3 Mouse HSC data

Finally, we apply our gene set analysis methods to an expression profiling study of HSCs from 30 BXD mice strains. We do not use the ANOVA model for differential expression because we do not have a definitive grouping of the individuals based on any one phenotype. Instead, we look for evidence of differential expression using the variance of the corrected expression levels. To compare our approach to the GSEA method, we compute the correlation between each gene and a known blood phenotype, the cycling of HSC after 7 days using the Fisher's *Z*-score. We use the gene sets from Tian *et al*. (2005).

Table 3 shows representative gene sets from the top 25 scoring clusters. Many of these gene sets represent housekeeping genes but we note some interesting candidates. The hypoxia signaling pathway, which includes genes such as HIF-1, a well known transcription factor that targets growth factors such as VEGF and EPO shows evidence of differential expression ($P < 0.001$) and tight regulation ($P < 0.006$). The growth factors in turn promote HSC development

Table 2. Significant Alzheimer's gene sets

Gene Set	DE P	DE Rank	TR P	TR Rank	Avg Rank	GSEA P	GSEA Rank	Set Size
KEGG:Oxidative_phosphorylation	1.47E-12	1	5.00E-04	10.5	*5.75*	1.04E-40	1	132
GO:0005740:mitochondrial membrane	1.76E-08	15	3.50E-03	36	*25.5*	4.12E-20	13	126
KEGG:Cholera_-_Infection	2.38E-04	54	5.00E-04	10.5	*32.3*	7.63E-08	101	88
GO:0015078:hydrogen ion transporter activity	**7.99E-12**	**2**	**1.20E-02**	**74**	***38***	**2.28E-37**	**5**	**131**
GO:0015399:primary active transporter activity	1.10E-11	3	1.20E-02	74	*38.5*	6.31E-39	2	170
GO:0006096:glycolysis	1.55E-03	67	5.00E-04	10.5	*38.8*	9.00E-12	40	44
GO:0005386:carrier activity	1.91E-11	4	1.20E-02	74	*39*	9.75E-30	8	274
GO:0015077:monovalent inorganic cation transporter activity	**8.37E-11**	**5**	**1.20E-02**	**74**	***39.5***	**1.86E-37**	**4**	**138**
GO:0008324:cation transporter activity	**2.72E-09**	**8**	**1.20E-02**	**74**	***41***	**3.30E-35**	**6**	**321**
GO:0003774:motor activity	2.24E-03	72	5.00E-04	10.5	*41.3*	5.20E-04	197	89
GO:0019866:inner membrane	4.89E-09	9	1.20E-02	74	*41.5*	1.41E-26	9	109
BioCyc:gluconeogenesis	4.67E-03	82	5.00E-04	10.5	*46.3*	3.27E-08	92	27
GO:0016675:oxidoreductase activity, acting on heme group of donors	4.74E-03	84.5	5.00E-04	10.5	*47.5*	1.10E-08	86.5	26
GO:0016676:oxidoreductase activity, acting on heme group of donors, oxygen as acceptor	4.74E-03	84.5	5.00E-04	10.5	*47.5*	1.10E-08	86.5	26
GO:0006091:generation of precursor metabolites and energy	2.40E-07	23	1.20E-02	74	*48.5*	2.42E-26	10	394
GO:0019320:hexose catabolism	5.29E-03	88	5.00E-04	10.5	*49.3*	7.59E-08	99	53
GO:0051258:protein polymerization	5.77E-03	90	5.00E-04	10.5	*50.3*	3.77E-06	133	39
GO:0015075:ion transporter activity	**1.29E-09**	**6**	**1.30E-02**	**99**	***52.5***	**1.34E-32**	**7**	**377**
BioCyc:aerobic respiration -- electron donors reaction list	5.44E-04	57	1.00E-02	48.5	*52.8*	1.92E-15	18	42
GO:0005507:copper ion binding	1.32E-03	63	8.50E-03	44	*53.5*	5.00E-02	384	36
GO:0019001:guanyl nucleotide binding	5.52E-04	59	1.05E-02	51	*55*	1.36E-06	124	251
GO:0015002:heme-copper terminal oxidase activity	4.74E-03	84.5	1.50E-03	27.5	*56*	1.10E-08	86.5	26
GO:0004129:cytochrome-c oxidase activity	4.74E-03	84.5	1.50E-03	27.5	*56*	1.10E-08	86.5	26
humanpaths: G-Proteins / Signaling Molecules	1.06E-02	105	5.00E-04	10.5	*57.8*	7.90E-06	141	125

Table 3. Significant mouse hematopoietic stem cell gene sets

Gene Set	DE P	DE Rank	TR P	TR Rank	Avg Rank	GSEA P	GSEA Rank	Set Size
KEGG:Ribosome	1.00E-20	1	5.00E-04	12	*6.5*	1.38E-04	49	71
GO:0030529:ribonucleoprotein complex	1.17E-14	2	5.00E-04	12	*7*	2.82E-07	13	235
GO:0003735:structural constituent of ribosome	7.85E-14	3	5.00E-04	12	*7.5*	2.28E-06	14	137
GO:0005840:ribosome	4.32E-13	4	5.00E-04	12	*8*	1.35E-07	11	120
GO:0005830:cytosolic ribosome (sensu Eukaryota)	1.62E-10	5	5.00E-04	12	*8.5*	6.60E-02	247	42
GO:0042254:ribosome biogenesis and assembly	6.58E-08	7	5.00E-04	12	*9.5*	1.00E-02	139	70
GO:0005198:structural molecule activity	2.74E-07	8	5.00E-04	12	*10*	5.00E-01	579	289
GO:0007028:cytoplasm organization and biogenesis	7.01E-07	9	5.00E-04	12	*10.5*	2.40E-02	186	78
GO:0007046:ribosome biogenesis	8.35E-07	10	5.00E-04	12	*11*	2.60E-02	191	63
GO:0005829:cytosol	1.68E-06	11	5.00E-04	12	*11.5*	2.40E-02	186	199
GO:0006412:protein biosynthesis	7.67E-06	12	5.00E-04	12	*12*	3.20E-02	203	307
GO:0009059:macromolecule biosynthesis	7.81E-05	15	5.00E-04	12	*13.5*	3.00E-02	200	343
GO:0006996:organelle organization and biogenesis	0.00014	21	5.00E-04	12	*16.5*	8.80E-01	767	392
GO:0044249:cellular biosynthesis	0.00155	37	5.00E-04	12	*24.5*	4.00E-03	104	493
GO:0016043:cell organization and biogenesis	0.0017	38	5.00E-04	12	*25*	7.94E-01	723	457
GO:0000785:chromatin	0.00184	40	5.00E-04	12	*26*	3.86E-01	512	63
GO:0006323:DNA packaging	4.75E-05	13	0.0045	45	*29*	3.76E-01	503	94
GO:0003723:RNA binding	2.76E-10	6	0.0115	55.5	*30.8*	1.04E-01	283	284
GO:0005694:chromosome	0.00291	47	0.001	27	*37*	6.12E-01	624	118
GO:0051258:protein polymerization	8.80E-05	16	0.017	64.5	*40.3*	3.30E-01	481	24
GO:0006457:protein folding	0.00194	41	0.0035	40	*40.5*	1.00E-02	139	106
mousepaths: Hypoxia Signaling Pathway	0.00145	36	0.0055	47.5	*41.8*	7.40E-01	691	65
GO:0006325:establishment and/or maintenance of chromatin architecture	0.00011	19	0.0185	66.5	*42.8*	7.64E-01	702	86
KEGG:Gap_junction	0.00856	74	5.00E-04	12	*43*	4.44E-01	546	56
GO:0015399:primary active transporter activity	0.00809	71	0.004	43	*57*	1.98E-09	6	121
GO:0015077:monovalent inorganic cation transporter activity	0.03021	126	0.001	27	*76.5*	1.48E-08	9	103

and differentiation. Although GSEA does not find the hypoxia signaling pathway ($P < 0.688$), it is able to find the KEGG insulin signaling pathway which contains the SHIP inositol phosphatase, *Inpp5d*, an important negative regulator of cytokine and immune receptor signaling. Absence of SHIP has been shown to affect the homeostasis and regeneration of murine HSCs (Helgason *et al*., 2003). Interestingly, insulin, IGF-1 and hypoxia have all been shown to stimulate HIF-1 leading one to speculate as to the complex nature of the regulatory network (Fukuda *et al*., 2002). We also find the KEGG gap junctions pathway Figure 1b (ranked #25) which contains connexin 43, a gene that is involved in multiconnexin-expressing stromal support of hematopoietic progenitors and stem cells (Cancelas *et al*., 2000). The information exchange mediated by stromal and HSC interaction has been implicated in stem cell differentiation and development (Orlic *et al*., 2001; Rosendaal *et al*., 1994).

Both GSEA and our method find two gene sets related to active transport (GO:0015399 and GO:0015077) which are not known to be related to any blood phenotypes. When we perform the GSEA analysis using two other phenotypes (stem cell frequency and concurrence of peripheral blood CD4%), we find some interesting results. Using stem cell frequency as the phenotype, these two gene sets are ranked 6 and 9 having *P*-values of $P < 1.98E\text{-}09$ and $P < 1.48E\text{-}08$, close to those found using the HSC turnover phenotype. However, when we use the concurrence of peripheral blood phenotype, the two gene sets are ranked 554 and 451 with *P*-values of $P < 0.69$ and $P < 0.40$. From the GSEA results, we see that these two gene sets are correlated with some phenotypes but not others. This result supports what we were able to find using expression data alone.

4 DISCUSSION

We have introduced a new method for analyzing gene sets based on the intuitive notions of tight regulation and differential expression. We first presented an approach to look for tightly regulated gene sets using correlation statistics. By considering the correlation structure of gene sets, we are restricting ourselves to only those gene sets that have pathway-like properties. This allows us to use gene sets from sources such as Gene Ontology, where biological pathways are not well represented, to their full potential. This is evident when we analyzed the Alzheimer's dataset. We found that while many gene sets associated with calcium ion channels are differentially expressed but not tightly regulated, other gene sets associated with the ATP-dependent calcium transporter are both differentially expressed and tightly regulated. This result suggests a regulatory mechanism at the level of the transporter and not the channel proteins.

We have also presented a more general framework for finding differentially expressed gene sets which is more appropriate for genome wide expression data over genetically similar individuals or strains of model organisms where clear phenotypic classes are absent and regulatory pattern differences are due to genetic variation. Our approach allows us to detect these subtle differences in expression of individuals which can then be used to look for associated phenotypes (Ghazalpour *et al*., 2005) as well as eQTLs in genetic variation studies (Bystrykh *et al*., 2005). In murine HSCs, we were able to find two active transport gene sets independent of phenotype information. Their correlation with phenotypes were then verified using a popular gene set based approach.

ACKNOWLEDGEMENTS

C.Y. and E.E. are partially supported by National Science Foundation Grant No. 0513612. Part of this investigation was supported using the computing facility made possible by the Research Facilities Improvement Program Grant Number C06 RR017588 awarded to the Whitaker Biomedical Engineering Institute, and the Biomedical Technology Resource Centers Program Grant Number P41 RR08605 awarded to the National Biomedical Computation Resource, UCSD, from the National Center for Research Resources, National Institutes of Health. Additional computational resources were provided by the California Institute of Telecommunications and Information Technology (Calit2).

REFERENCES

Berriz,G.F. *et al*. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.

Blalock,E.M. *et al*. (2004) Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl Acad. Sci. USA*, **101**, 2173–2178.

Bystrykh,L. *et al*. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet*., **37**, 225–232.

Cancelas,J.A. *et al*. (2000) Connexin-43 gap junctions are involved in multiconnexin-expressing stromal support of hemopoietic progenitors and stem cells. *Blood*, **96**, 498–505.

Cui,X. and Churchill,G. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*., **4**, 210.

Dahlquist,K.D. *et al*. (2002) Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet*., **31**, 19–20.

Draghici,S. *et al*. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*., **31**, 3775–3781.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fukuda,R. *et al*. (2002) Insulin-like growth factor 1 induces hypoxia-inducible factor 1-mediated vascular endothelial growth factor expression, which is dependent on MAP kinase and phosphatidylinositol 3-kinase signaling in colon cancer cells. *J. Biol. Chem*., **277**, 38205–38211.

Ghazalpour,A. *et al*. (2005) Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol*., **6**, R59.

Helgason,C.D. *et al*. (2003) Homeostasis and regeneration of the hematopoietic stem cell pool are altered in SHIP-deficient mice. *Blood*, **102**, 3541–3547.

Hinerfeld,D. *et al*. (2004) Endogenous mitochondrial oxidative stress: neurodegeneration, proteomic analysis, specific respiratory chain defects, and efficacious antioxidant therapy in superoxide dismutase 2 null mice. *J. Neurochem*., **88**, 657–667.

Hosack,D. *et al*. (2003) Identifying biological themes within lists of genes with ease. *Genome Biol*., **4**, R70. Available at http://genomebiology.com/2003/4/6/P4.

Mootha,V.K. *et al*. (2003) Pgc-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*., **34**, *267–273*.

Morley,M. *et al*. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Orlic,D. *et al*. (2001) Bone marrow cells regenerate infarcted myocardium. *Nature*, **410**, 701–705.

Pavlidis,P. *et al*. (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res*., **29**, 1213–1222.

Perou,C.M. *et al*. (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

R Development Core Team (2005) R Foundation for Statistical Computing, Vienna, Austria*R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Randle,P.J. *et al*. (1964) Regulation of glucose uptake by muscle. 8. effects of fatty acids, ketone bodies and pyruvate, and of alloxan-diabetes and starvation, on the uptake and metabolic fate of glucose in rat heart and diaphragm muscles. *Biochem. J*., **93**, 652–665.

Rosendaal,M. *et al*. (1994) Up-regulation of the connexin43+ gap junction network in haemopoietic tissue before the growth of stem cells. *J. Cell Sci*., **107**, 29–37.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Subramanian,A. *et al*. (2005) From the cover: gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tian,L. *et al*. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.

Zeeberg,B. *et al*. (2003) Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*., **4**, R28.

Zhong,S. *et al*. (2003) ChipInfo: software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Res*., **31**, 3483–3486.