Analysis of Sample Set Enrichment Scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles

Elena Edelman [†,‡], Alessandro Porrello[†,*,⋆,⊙], Justin Guinney[†,‡], Bala Balakumaran[†,*,⋆], Andrea Bild[†,*], Phillip G. Febbo[†,*,⋆], Sayan Mukherjee[†,⋄,◊]

[†] Institute for Genome Sciences & Policy, [‡] Computational Biology and Bioinformatics Program, [*] Department of Medicine, [⋆] Department of Molecular Genetics and Microbiology, [⋄] Institute of Statistics and Decision Sciences, [◊] Department of Computer Science, Duke University, Durham NC 27708, USA. [⊙] Molecular Oncogenesis Laboratory, Regina Elena Cancer Institute, Via Delle Messi D'Oro 156, Rome, 00158, Italy.

ABSTRACT

Motivation: Gene expression profiling experiments in cell lines and animal models characterized by specific genetic or molecular perturbations have yielded sets of genes annotated by the perturbation. These gene sets can serve as a reference base for interrogating other expression datasets. For example, a new dataset in which a specific pathway gene set appears to be enriched, in terms of multiple genes in that set evidencing expression changes, can then be annotated by that reference pathway. We introduce in this paper a formal statistical method to measure the enrichment of each sample in an expression dataset. This allows us to assay the natural variation of pathway activity in observed gene expression data sets from clinical cancer and other studies.

Results: Validation of the method and illustrations of biological insights gleaned are demonstrated on cell line data, mouse models, and cancer-related datasets. Using oncogenic pathway signatures, we show that gene sets built from a model system are indeed enriched in the model system. We employ ASSESS for the use of molecular classification by pathways. This provides an accurate classifier that can be interpreted at the level of pathways instead of individual genes. Finally, ASSESS can be used for cross-platform expression models where data on the same type of cancer are integrated over different platforms into a space of enrichment scores.

Availability: Versions are available in Octave and Java (with a graphical user interface). Software can be downloaded at http://people.genome.duke.edu/assess.

Contact: sayan@stat.duke.edu

1 INTRODUCTION

Gene expression profiling experiments have been conducted on a wide variety of cell lines and animal models with the goal of characterizing genes sets whose expression patterns characterize specific genetic or molecular perturbations. These gene sets contain candidate players in pathways, or sub-pathways, that are "annotated" by the experimental perturbation. The fundamental idea in pathway based analysis approaches (Huang *et al.*, 2003; Black *et al.*, 2003; Mootha *et al.*, 2003; Sweet-Cordero *et al.*, 2005; Alvarez *et al.*, 2005; Febbo *et al.*, 2005; Subramanian *et al.*, 2005) is that such

a gene set serves as a reference base for interrogating other expression data sets. A new data set in which a specific pathway gene set appears to be enriched, in terms of multiple genes in that set evidencing expression changes can then be annotated by that reference pathway. An analogy can be made here with sequence annotation in a BLAST search: sets of experimentally derived pathways serve as annotation reference sets for future experiments in the same way that annotated sequences serve as references in a sequence search. Statistical methods are needed and have been developed (Subramanian *et al.*, 2005; Barry *et al.*, 2005; Kim and Volsky, 2005; Tomfohr *et al.*, 2005) to define computational tools for such expression-based pathway annotation. Two of these methods, GSEA (Subramanian *et al.*, 2005) and SAFE (Barry *et al.*, 2005), use nonparametric statistics to provide formal statistical evaluation, and confidence assessments, for annotation of an expression data set by measuring the overlap of significantly perturbed genes with those in each pathway in a database of pathways. Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005) has been successfully applied in many basic science and clinical studies (Mootha *et al.*, 2003; Sweet-Cordero *et al.*, 2005; Alvarez *et al.*, 2005; Febbo *et al.*, 2005; Subramanian *et al.*, 2005; Bild *et al.*, 2006), including pathway deregulation in cancer genomics. A fundamental shortcoming of GSEA and other methods (Barry *et al.*, 2005; Kim and Volsky, 2005; Tomfohr *et al.*, 2005) is that they do not characterize the variation in enrichment over individual samples in the data set.

If the enrichment of each sample in an expression data set can be annotated, then one can assay the natural variation of pathway activity in observed gene expression data sets from clinical cancer and other studies. The ability to assay pathway variation in samples allows the implementation of a general methodology to dissect tumor samples in terms of oncogenic pathways. The logic behind this methodology is to develop gene expression "signatures" of oncogenic pathways from model systems and then use these "signatures" to annotate human tumors in terms of the deregulation of oncogenic pathways (Bild *et al.*, 2006).

In this paper we introduce a statistical method that allows us to assay pathway variation, Analysis of Sample Set Enrichment Scores (ASSESS). Given gene sets defined by prior biological knowledge or genes co-expressed in an experiment with a specific genetic or

molecular perturbation, and a data set of expression profiles from samples belonging to two classes, ASSESS provides: a measure of the enrichment of each gene set in each sample and a confidence assessment. This extends the methodology developed in GSEA and SAFE to annotate individual samples.

A family of methods for pathway annotation was developed and used to measure pathway deregulation in breast cancer and lung cancer (Huang *et al.*, 2003; Black *et al.*, 2003; Bild *et al.*, 2006). The approach involved: (a) building statistical models of pathway deregulation from cell lines where recombinant adenoviruses were used to express oncogenic activities corresponding to pathway deregulation, (b) applying the models to each sample in a data set of tumors and estimating the probability of deregulation of the pathways. The main drawback of this methodology is that cell line perturbation data as well as tumor data are required for the analysis. For ASSESS, only the list of genes characterizing the pathway deregulation is required, the entire model and cell line data is not needed. This provides a great advantage when the gene sets are determined by literature review or a non-expression based assay, such as immunohistochemical, for which building an accurate model subsequently applicable to expression data is a very difficult challenge.

2 RESULTS

2.1 Analysis of Sample Set Enrichment Scores

ASSESS is an annotation methodology that takes as inputs:

1. Genome-wide expression profiles consisting of p genes and n samples with each sample corresponding to one of two classes, C_1, C_2. The expression of the j-th gene in the i-th sample is x_j^i;

2. A database of m gene sets $\Gamma = \{\gamma_1, ... \gamma_m\}$ where each gene set γ_k is a list of genes (a subset of the p genes in the data set) belonging to a pathway or other functional or structural category;

3. A ranking procedure and correlation statistic that takes the expression data set and labels as inputs and produces correlation statistics for each sample that reflects the correlation of the p genes in that sample with respect to the the distribution of expression in the two classes. The correlation statistics for the i-th sample would be $\mathbf{c}_i = \{c_1^i, ..., c_p^i\}$ where c_j^i is calculated by any likelihood ratio statistic for measuring the correlation of a sample to one class rather than the other:

$$c_j^i = \log \left(\frac{\mathbb{P}(x_j^i \in C_1 | x_j^i, \text{data})}{\mathbb{P}(x_j^i \in C_2 | x_j^i, \text{data})} \right).$$

An example of a parametric and nonparametric correlation statistic is described in detail in the following sections;

and produces as outputs:

1. An enrichment score for each sample in the data set with respect to each gene set in the database. ES_i^k corresponds to the enrichment of the i-th sample with respect to the k-th gene set;

2. A measure of confidence for each enrichment score is given by a p-value with multiplicity taken into account by Family-wise error rate (FWER) p-values and False Discovery rate (FDR) q-values.

Given the correlation statistics for the i-th sample, $\mathbf{c}_i = \{c_1^i, ..., c_p^i\}$, and a gene set γ_k, we construct the following discrete random walk over the indices of the rank-ordered correlation statistic

$$\nu(\ell) = \frac{\sum_{j=1}^{\ell} |c_{(j)}|^\tau I(g_{(j)} \in \gamma_k)}{\sum_{j=1}^{p} |c_{(j)}|^\tau I(g_{(j)} \in \gamma_k)} - \frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma_k)}{p - |\gamma_k|}, \quad (1)$$

where $c_{(j)}$ is the rank-ordered correlation statistic, τ is a parameter (in general $\tau = 1$), γ_k is the k-th gene set, $I(g_{(j)} \in \gamma_k)$ is the indicator function on whether the j-th gene (the gene corresponding to the j-th ranked correlation statistic) is in gene set γ_k, $|\gamma_k|$ is the number of genes in the k-th gene set, and p is the number of genes in the data set. The enrichment statistic for the i-th sample with respect to the k-th gene set is the maximum deviation of the random walk from zero

$$ES_i^k = \nu[\arg \max_{\ell=1,..,p} |\nu(\ell)|]. \quad (2)$$

The random walk is a tied-down Brownian bridge process and the deviation from zero is very closely related to the classical Kolmogorov-Smirnov statistic (Feller, 1971). There are simpler ways to define the enrichment score such as taking the average rank of the genes in the gene set from the rank-ordered list of genes. However, using a random walk is advantageous because it allows one to see how the genes in the set are distributed in the rank-ordered list. The random walk could alternatively be solved by ranking $|c_{(j)}|$ rather than $c_{(j)}$. While in certain situations a more extreme enrichment score may be sacrificed, this ranking will not allow for the access to the additional information of which genes in the set are up or down regulated. Therefore we choose to calculate the ES by a random walk using values ranked by $c_{(j)}$, and suggest gene sets be constructed to capture genes correlated with either over expression or under expression in a class but not both.

To measure significance we assume under the null hypothesis that the labels are exchangeable and therefore we can compute the null distribution by permuting labels, ranking the genes according to the recomputed statistic $c_{(j)}^\pi$, and computing the "random" enrichment statistic $ES_i^k(\pi)$. This is done over many label permutations, $\pi = 1, ..., \Pi$. The p-value is computed by comparing the enrichment score to the empirical distribution generated from $\{ES_i^k(\pi)\}_{\pi=1}^{\Pi}$. Correction for multiple hypothesis testing is preformed in the same manner as in GSEA and is addressed via FWER p-values or FDR q-values (see (Subramanian *et al.*, 2005) for details). Overlapping gene sets do not influence the calculation of q-values.

The key technical innovation in extending methods such as GSEA or SAFE to provide enrichment scores for individual samples is producing a correlation statistic and subsequent rankings that model how representative each gene for a given sample is with respect to the two classes. The ranking should reflect the natural variation of how each sample is correlated with class labels. We introduce two correlation statistics which reflect this variation: (1) based on a simple parametric normal model, (2) based on a nonparametric random walk model. Both methods preform well, as likely would other well-defined likelihood ratios such as a binary regression model. We choose to use the nonparametric model on the data in this paper as this method is a novel means for calculating class membership likelihoods. However, the parametric method as well as other methods not discussed here give comparable results.

2.1.1 Parametric model The parametric model assumes that the expression of a given gene can be modeled by a mixture of two normal distributions corresponding to the two classes. The mean and standard deviations are computed from the data

$$\hat{\mu}_{j1} = \frac{1}{n_1} \sum_{i \in C_1} x_j^i, \qquad \hat{\mu}_{j2} = \frac{1}{n_2} \sum_{i \in C_2} x_j^i,$$

$$\hat{\sigma}_{j1}^2 = \frac{1}{n_1} \sum_{i \in C_1} (x_j^i - \hat{\mu}_{j1})^2, \qquad \hat{\sigma}_{j2}^2 = \frac{1}{n_2} \sum_{i \in C_2} (x_j^i - \hat{\mu}_{j2})^2,$$

where n_1 and n_2 are the number of samples in class 1 and 2. The expression of the j-th gene is modeled as $\mathbf{N}(\hat{\mu}_{j1}, \hat{\sigma}_{j1})$ or $\mathbf{N}(\hat{\mu}_{j2}, \hat{\sigma}_{j2})$ depending on whether the sample belongs to class 1 or 2. We define the class membership likelihood of expression x_j^i from the models of classes 1 and 2 as p_{j1} and p_{j2} respectively.

$$p_{j1} = \mathbb{P}(\xi \geq x_j^i | \xi \sim \mathbf{N}(\hat{\mu}_{j1}, \hat{\sigma}_{j1})), \quad \text{if } x_j^i \geq \hat{\mu}_{j1},$$
$$p_{j1} = \mathbb{P}(\xi < x_j^i | \xi \sim \mathbf{N}(\hat{\mu}_{j1}, \hat{\sigma}_{j1})), \quad \text{if } x_j^i < \hat{\mu}_{j1},$$
$$p_{j2} = \mathbb{P}(\xi \geq x_j^i | \xi \sim \mathbf{N}(\hat{\mu}_{j2}, \hat{\sigma}_{j2})), \quad \text{if } x_j^i \geq \hat{\mu}_{j2},$$
$$p_{j2} = \mathbb{P}(\xi < x_j^i | \xi \sim \mathbf{N}(\hat{\mu}_{j2}, \hat{\sigma}_{j2})), \quad \text{if } x_j^i < \hat{\mu}_{j2}.$$

We use the distribution function rather than the density because there is a very natural directionality assumption in this model in that if the Gaussians are well separated then the deeper inside the respective class a point x resides the higher should be the membership probability. We then use the log-likelihood ratio as the correlation statistic. Given expression, x_j^i, of the j-th gene of the i-th sample the correlation statistic is computed as:

$$c_j^i = \log\left(\frac{p_{j1}}{p_{j2}}\right), \quad \text{if } \hat{\mu}_{j1} \geq \hat{\mu}_{j2}$$

$$c_j^i = \log\left(\frac{p_{j2}}{p_{j1}}\right), \quad \text{otherwise.}$$

Thus, genes are ranked based upon the differential probability of their membership in either class and because of this, genes are ranked as a continuum from those with the greatest probability of belonging to class 1 ranked at the top and genes with the greatest probability of belonging to class 2 near the bottom. As most genes will have limited differential expression between the two classes, these genes will have similar probabilities of belonging to either group and the log-likelihood ratio will be near zero.

2.1.2 Nonparametric model The assumption of normality in the parametric model is often inappropriate for expression data. For this reason, a nonparametric model to compute class membership likelihoods is used in most applications. The class membership likelihoods are computed based upon absorption probabilities of a Brownian motion (random walk) (see Figure 1 for an illustration of the model).

We first estimate the densities of the j-th gene for the two classes, $\hat{p}_{j1}(x)$ and $\hat{p}_{j2}(x)$, using a Parzen estimator (Vapnik, 1998) with a Gaussian kernel:

$$\hat{p}_{j1}(x) = \frac{1}{n_1 \sigma_{j1} \sqrt{2\pi}} \sum_{i \in C_1} e^{-|x_j^i - x|^2 / 2\sigma_{j1}^2},$$

$$\hat{p}_{j2}(x) = \frac{1}{n_2 \sigma_{j2} \sqrt{2\pi}} \sum_{i \in C_2} e^{-|x_j^i - x|^2 / 2\sigma_{j2}^2},$$

where n_1 and n_2 are the number of samples in C_1 and C_2 and bandwidths σ_{j1} and σ_{j2} are set to the average distance between points of the j-th gene in C_1 and C_2 respectively. We define two points, e_{j1}^i and e_{j2}^i, as the left or right extremes of the random walk starting at x_j^i.

$$e_{j1}^i = \min_i\{x_j^i\} \text{ if } x_j^i < \hat{\mu}_{j1}, \qquad e_{j1}^i = \max_i\{x_j^i\} \text{ if } x_j^i \geq \hat{\mu}_{j1},$$

$$e_{j2}^i = \min_i\{x_j^i\} \text{ if } x_j^i < \hat{\mu}_{j2}, \qquad e_{j2}^i = \max_i\{x_j^i\} \text{ if } x_j^i \geq \hat{\mu}_{j2}.$$

The membership likelihood of expression x_j^i for class 1 and 2 is given by the absorption probability at the points e_{j1}^i and e_{j2}^i for a Brownian motion starting at x_j^i with initial conditions distributed as $\hat{p}_{j1}(x)$ and $\hat{p}_{j2}(x)$.

We again use the log-likelihood ratio as the correlation statistic. Given expression, x_j^i, of the j-th gene of the i-th sample the correlation statistic is computed as:

$$c_j^i = \log\left(\frac{\mathbb{P}(\text{absorption at } e_{j1}^i \text{ starting at } x_j^i | \hat{p}_{j1})}{\mathbb{P}(\text{absorption at } e_{j2}^i \text{ starting at } x_j^i | \hat{p}_{j2})}\right). \quad (3)$$

The absorption probabilities can be computed as the solution of the Dirichlet problem (Durrett, 1996) which for the Parzen estimators results in a weighted sum of error functions and exponentials (see methods section for the exact form and derivation). So the correlation statistics can be computed efficiently.
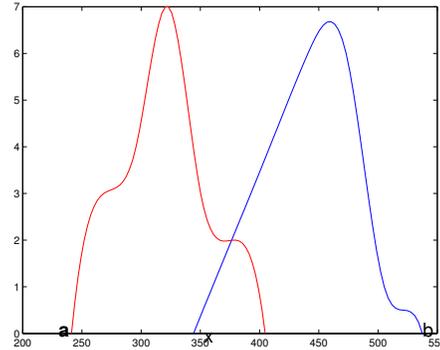
Fig. 1. The two classes' densities are displayed by the red and blue curves, p_r, p_b. Assume we have a diffusion process (random walk) starting at x. We compute the probability of absorption at the point b if the initial conditions are distributed as p_b, $\mathbb{P}(\text{absorption at b starting at x} | p_b)$. We also compute the probability of absorption at the point a if the initial conditions are distributed as p_r, $\mathbb{P}(\text{absorption at a starting at x} | p_r)$. These two probabilities serve as a measure that an individual sample belongs to one of the two distributions.

2.2 Validation on Model Systems

The objective of ASSESS is to annotate each sample in an expression data set in terms of a priori defined gene sets often constructed from model systems. In this section we validate the method by demonstrating that gene sets built from model systems or with known genetic perturbations are indeed enriched in gene expression data from the same model systems or related systems.

2.2.1 Mouse models In (Majumder *et al.*, 2004) transgenic mice were generated that developed a highly penetrant prostatic intraepithelial neoplasia (PIN) phenotype and expressed a constitutively active AKT1 gene in the ventral prostate of the mouse. This AKT-induced PIN phenotype can be reversed with treatment of RAD001, a mTOR inhibitor. The transgenic mice were split into two groups, with one group receiving RAD001 and the other a placebo. Tissue was taken from the prostate of both groups and DNA microarray analysis was preformed using the Affymetrix Murine U430A microarray. This resulted in two sets of expression data: samples treated with RAD001 ($n = 19$) and placebo ($n = 19$). These data sets were split into a training and test set. The training set consisted of the first 10 samples treated with RAD001 and the first 10 samples treated with the placebo. The test set was comprised of the complimentary samples. The training set was used to construct an AKT gene set using a logistic regression model (see methods section for details).

We applied ASSESS to the test set using the AKT gene set derived from the training data. The enrichment scores of the samples treated with RAD001 strongly indicate that genes in the AKT gene set were under expressed compared to the samples given placebo which showed enrichment in the gene set (see Figure 2). All samples were significantly enriched (p-value < 0.001).
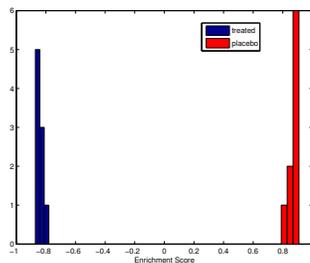
Fig. 2. A histogram of enrichment scores for the 9 treated and 9 untreated mouse prostate samples in the test data with respect to the AKT pathway gene set computed from the training data.

2.2.2 Cell culture models In (Bild *et al.*, 2006) human primary mammary epithelial cell cultures (HMECs) were used to develop a series of pathway signatures which were then used to assay pathway disregulation in non-small cell lung carcinoma (NSCLC). We use this data set to validate our method.

The data was generated by using recombinant adenoviruses to express specific oncogenes in an otherwise quiescent cell, thereby isolating the subsequent events as defined by the activation/deregulation of that single pathway. The cells were infected with adenovirus expressing either human c-Myc, activated H-Ras, human c-Src, human E2F3, or activated β-cantenin. RNA from these multiple independent infections, as well as from normal cells (with green florescent protein, GFP), was collected for DNA microarray analysis using the Affymetrix Human Genome U133 Plus 2.0 Array.

Given the independent replicates from the six conditions, the five perturbed pathways and the normal GFP cells, we split each condition into a train and test set. Thus given expression data from: 10 Myc, 10 Ras, 7 Src, 10 E2F3, 9 β-catenin, and 10 normal/GFP

samples we construct five training sets with the first half of the samples from each experimental data set along with the first 5 normal samples. Similarly, five independent test sets were constructed using the complimentary samples (the second half of samples in the six conditions). The training sets were used to construct gene sets for each of the five pathways, Myc, Ras, Src, E2F3, and β-catenin using a logistic regression model (see methods section for details).

ASSESS was applied to the five test sets to calculate enrichment with respect to the five gene sets computed from the training data.

Experiment	ES for GFP cells	ES for infected cells
BCAT	-0.87(±0.031)	0.88(±0.042)
E2F3	-0.97(±0.0069)	0.98(±0.0061)
MYC	-0.89(±0.018)	0.91(±0.067)
RAS	-0.96(±0.012)	0.91(±0.021)
SRC	-0.90(±0.016)	0.91(±0.022)

Table 1. Average enrichment scores(±sd) for the comparison of normal (GFP cells) to infected cells for the gene set built from the respective infected cell type.

2.2.3 Literature based models The approach developed in (Huang *et al.*, 2003; Black *et al.*, 2003; Bild *et al.*, 2006) of building statistical models of pathway deregulation in controlled experiments and then applying these to new data sets could have been used in the previous two examples. However, this approach requires that the cell line perturbation data as well as new data and that the data are on comparable platforms. This approach cannot be used for gene sets derived from literature whereas ASSESS is still applicable.

In (Subramanian *et al.*, 2005) a data set generated from mRNA expression from lymphoblastoid cell lines derived from 15 males and 17 females served as a validation set. We then sought to address the following: which gene sets were over expressed in males and which in females. Gene sets defined by cytogenetic bands and gene sets defined by pathway or functional properties were examined. As expected for males, chromosome Y as well as its two bands (Yp11 and Yq11) and a gene set corresponding to genes enriched in male reproductive tissue (testis) were overexpressed. For females two gene sets of genes that escape X-inactivation were overexpressed in addition to a gene set corresponding to genes enriched in female reproductive tissue (uterus). Genes on the X-chromosome would not be expected to be overexpressed due to dosage compensation by X-inactivation.

The enrichment of these seven gene sets with respect to the male and female samples in the lymphoblastoid cell lines is displayed in Figure 3. The male samples are clearly enriched with respect to: Y, Yq11, Yp11, and testis. The female samples are clearly enriched with respect to: the two escape of X-inactivation gene sets (X-inactivation and Willard X-inactivation) and the uterus gene set (labeled in the figure as Reproduction Genes). We used a Myc gene set as a control in that it is not expected to be enriched with respect to the male/female distinction and indeed this is the case.

We further illustrate the procedure by plotting the random walk (Equation (1)) for a male and female sample with respect to one of the escape from X-inactivation gene sets and a Myc gene set (see Figure 4). For a female sample with respect to this gene set, the
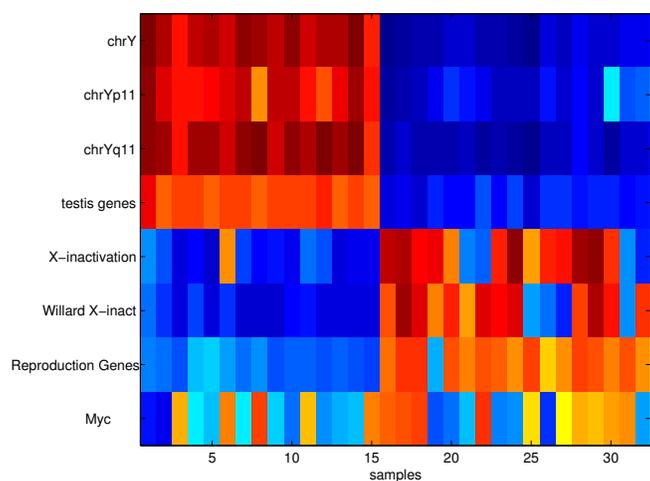
Fig. 3. Enrichment scores for the comparison of males to females in the 8 gene sets. The male samples $(1-15)$ show enrichment in the Y, Yq11, Yp11, and testis gene sets. The female samples $(16-32)$ show enrichment in the two escape of X-inactivation gene sets and the uterus gene set. The Myc pathway shows no differential expression between males and females, as expected.

random walk increases very rapidly initially indicating that genes escaping X-inactivation appear at the top of the list of genes ordered by correlation with the female phenotype. This results in a very positive enrichment score. For the male sample the random walk is basically a mirror image of the female case indicating that genes escaping X-inactivation appear at the end of a list of genes ordered by correlation with the male phenotype. This results in a very negative enrichment score. The third case is for a female sample with respect to the Myc gene set. In this case the genes in the gene set are randomly spread over the ranked list and so the enrichment score never deviates far from zero.

2.3 Classification and Clustering in the Space of Pathways

A very natural consequence of obtaining enrichment scores for each sample in the data set is that classification and clustering can now be performed in the space of gene sets rather than individual genes. Being able to interpret classification models using pathways offers an alternative and possibly more intuitive perspective than models using individual genes. Another aspect of building models in the space of pathways that was emphasized in (Bild *et al.*, 2006) is knowing which pathways are disregulated with respect to outcome and how this can help suggest targeted therapeutics.

2.3.1 Clustering In (Brunet *et al.*, 2004) a matrix factorization method (NMF) was applied to an expression data set with acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL) samples (Golub *et al.*, 1999). The matrix factorization allowed the clustering of the samples into subsets. A parameter in this clustering method is the number of subsets k. For this data set results with $k = 2, 3$ were computed. For the two cases the clusters comprised of $\{(25 \text{ ALL}), (11 \text{ AML}, 2 \text{ ALL})\}$ and $\{(8 \text{ ALL-T}), (17 \text{ ALL-B}), (11 \text{ AML}, 2 \text{ ALL-B})\}$, where ALL-T and ALL-B are two subtypes of ALL. We applied ASSESS to this
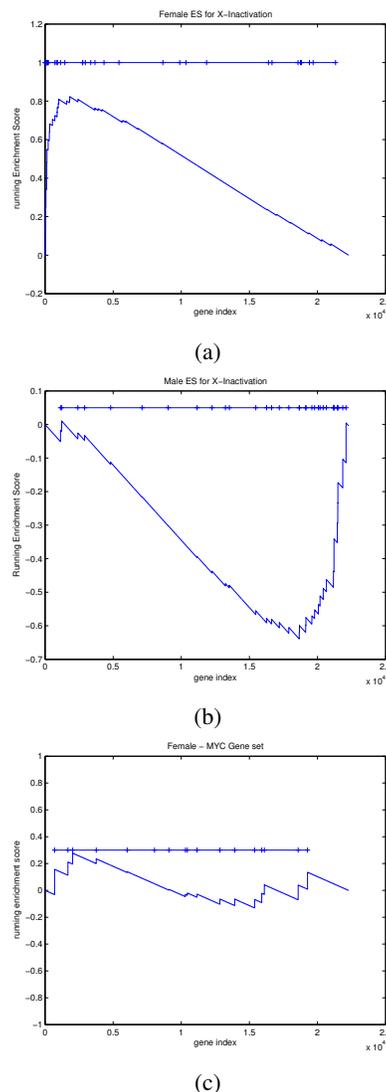
Fig. 4. Random walks. (a) The random walk for a female sample with respect to one of the escape from X-inactivation gene sets. The hatches of the top line indicate where the genes in the gene set fall with respect to the rank-ordering. (b) The random walk for a male sample with respect to one of the escape from X-inactivation gene sets. (c) The random walk for female sample with respect to a Myc gene set. This gene set is not significantly enriched and so the hatches appear randomly dispersed with respect to the rank-ordering.

leukemia data set using a database of 523 gene sets (Subramanian *et al.*, 2005). We then applied NMF to this space of enrichment scores and obtained identical results. The only difference is according to the measure of confidence developed in (Brunet *et al.*, 2004), as the clustering obtained from the enrichment space had greater confidence than that from the raw expression data. The result of the clustering and the factors are displayed in Figure 5.

2.3.2 Classification We examined six gene expression data sets for which single gene classification models have been built: (a) Gender – male vs. female (Subramanian *et al.*, 2005), (b) cDNA Lung cancer – squamous vs. adenocarcinoma (Garber *et al.*, 2001), (c)
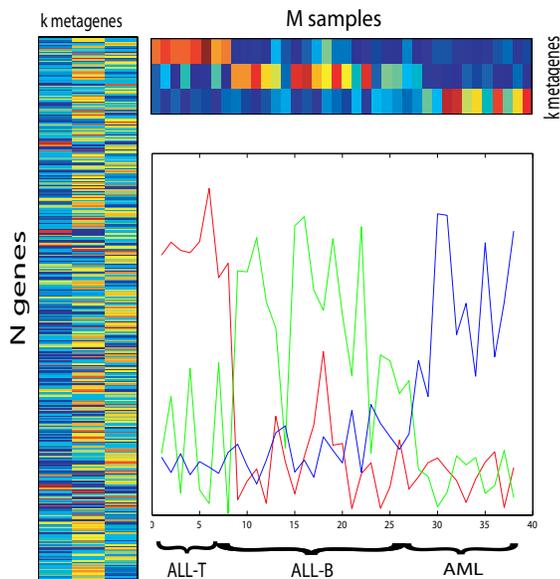
Fig. 5. The top and left figure are the left and right matrix factors for the matrix of enrichment scores in the Leukemia data with $k = 3$. The red line is a plot of the first metapathway over the data and this metapathway selects for the ALL-T samples. The green line is the second metapathway and it selects for the ALL-B samples. The blue line is a plot of the third metapathway which selects for the AML samples.

oligonucleotide Lung cancer – squamous vs. adenocarcinoma (Potti *et al.*, 2006), (d) Medulloblastoma – survival vs. failure (Pomeroy *et al.*, 2002), (e) Prostate cancer – recurrence vs. nonrecurrence (Glinsky *et al.*, 2004), and (f) Leukemia – AML vs. ALL (Golub *et al.*, 1999).

We applied the classification using enrichment scores procedure outlined in the methods section to compute classification accuracy on these six data sets (see Table 2). For all the data sets except for the Leukemia data set the leave-one-out method was used (Algorithm 1). For the Leukemia data set the train-test procedure was used (Algorithm 2) with the train-test split outlined in (Golub *et al.*, 1999). The classification accuracy was comparable or better than that for single gene classifiers for all the data sets except for the Leukemia data.

The pathways associated with recurrent prostate cancer tumors supports ASSESS's ability to both accurately predict outcome as well as provide biological insight. Both AKT and PTEN gene sets were found to have increased coordinate expression in samples of recurrent prostate cancer. PTEN loss is one of the most common genetic alterations seen in advanced prostate cancer resulting in activation of the PI3K-AKT pathway. Activation of this pathway is known to occur at a greater frequency in advanced prostate cancer and has prognostic significance. A "TERT-up" gene set was similarly found to be associated with recurrent prostate cancer. An essential requirement for tumor progression is avoidance of cellular senescence, telomerase restores chromosomal telomeres and is associated with the development of prostate cancer. Finally, another interesting observation is the presence of the "DNA damage

signaling" and "Cell cycle checkpoint" pathways both representing common cellular processes disregulated in aggressive cancer.

Classes	Accuracy
Gender: males vs. Females	94%
Lung Cancer(cDNA): Adeno vs. Squamous	91%
Lung Cancer(oglio): Adeno vs. Squamous	84%
Medulloblastoma: survival vs. failure	72%
Prostate: recurrent vs. nonrecurrent	73%
Leukemia: ALL vs. AML	85%

Table 2. Classification accuracy for six data sets building classification models in the space of enrichment scores.

2.4 Cross-platform Expression Models

DNA microarray studies have been carried out on a variety of platforms for the same case-control experiment, for example both cDNA microarrays and oligonucleotide microarrays are popular in cancer genomics. The integration of data across platforms is appealing for a variety of reasons: increasing the sample size of the data, allowing for interstudy validation, mitigating platform based biases, and mitigating study based biases.

Building a model from raw expression data from one platform and applying the model to data from another platform directly will not work since the expression data from the two platforms have different distributions. One approach to normalize between the platforms is to use median rank scores and and quantile discretization to map the data to a common space and then build a classification model in this space (Warnat *et al.*, 2005).

We advocate an alternative approach of applying ASSESS to expression data to map the data into the space of enrichment scores for pathways and then building models in this space. Methods such as median rank scores are no longer needed as enrichment scores between platforms are numerically comparable. There are several advantages to this approach: (1) the need to map genes using Uni-Gene ids is avoided; (2) the problem of multiple probe mappings between platforms is avoided; (3) gene sets defined separately by probes specific to each platform can be used; (4) the enrichment statistic is much more robust than the rank of a single gene so the loss of genes between platforms is mitigated; (5) interpreting results on the level of pathways instead of single genes is appealing.

We first applied this approach to two prostate cancer studies (Welsh *et al.*, 2001; Dhanasekaran *et al.*, 2001). The two platforms for the studies were cDNA microarrays (Dhanasekaran *et al.*, 2001) and Affymetrix oligonucleotide microarrays (Welsh *et al.*, 2001). The cDNA data set contained 53 samples of which 34 were tumors and 19 were normal. The oligo data set contained 33 samples of which 24 were tumors and 9 were normal. The catalog of human functional gene sets comprised of 433 sets annotated for both platforms (Subramanian *et al.*, 2005) was used as the gene set. The error rate for using the cDNA and oligo data sets as train-test sets respectively is reported in Table 3, as is the error rate for a leave-one-out procedure using all the cDNA and oligo samples (see methods section for details on on both test-train and leave-one-out classification using gene sets). We compare these results with the leave-one-out error computed on the individual data sets (see Table 3).

We next applied this approach to two lung cancer studies (Garber *et al.*, 2001; Potti *et al.*, 2006). The two studies involved the same platforms as the prostate example. The cDNA data set contained 55 samples of which 38 were adenocarcinomas and 17 were squamous cell lung carcinomas (Garber *et al.*, 2001). The oligo data set contained 93 samples of which 45 were adenocarcinomas and 48 were squamous cell lung carcinomas (Potti *et al.*, 2006). The same catalog of gene sets as used in the prostate example was used. The error rates for the cross-platform predictions as well as predictions within the individual data sets are summarized in Table 3.

	cDNA LOO	oligo LOO	train-test	combined LOO
prostate T/N	85.7%	76.5%	(cDNA-oligo) 73.5%	80.7%
lung A/S	88.0%	90.9%	(oligo-cDNA) 78.2%	88.5%

Table 3. Classification accuracy for cross-platform models for the prostate and lung cancer data sets.

3 METHODS

3.1 Gene set construction

Given an expression data set with two class labels, we use a linear logistic regression model with regularization or shrinkage (Hastie *et al.*, 2000) to construct gene sets. We define the expression data as a matrix x_j^i with $i = 1, ..., n$ (the number of samples) and $j = 1, ..., p$ (the number of genes), the i-th sample is designated as x_i, and the class labels as $y \in \{-1, 1\}$. The logistic regression model with regularization involves solving the following optimization problem

$$\arg\min_{w,b} \left[\frac{1}{n} \sum_{i=1}^n \log(1 + e^{-(y_i(w \cdot x_i + b))}) + \lambda \|w\|^2 \right],$$

where λ is a model parameter that needs to be set.

Solving the above optimization problem results in a vector $\hat{w}$ and the absolute magnitude of the elements of the vector correspond to the relevance of a gene or feature. For the HMEC data sets and the AKT data set, genes corresponding to 50 elements of $\hat{w}$ most correlated with the perturbation phenotype were used to construct the gene sets. In both algorithms 1 and 2 genes corresponding to the top and bottom 50 elements of $\hat{w}$ were used.

3.2 Classification and Gene Set Selection

Classification using enrichment scores was applied in two settings: a train-test setting and a leave-one-out cross-validation setting. The leave-one-out setting was used for all the data sets except the leukemia data set for which we used the test-train setting. The test-train setting is a simple generalization of the leave-one-out setting.

3.2.1 Leave-one-out setting: Given data set x_j^i of gene expression for $j = 1, .., p$ genes and $i = 1, ..., n$ samples where the i-th column of the matrix X correspond to the i-th sample, labels $(y_i)_{i=1}^n$, and gene sets $\Gamma = \{\gamma_1, ..., \gamma_m\}$ the leave-one-out method outlined in Algorithm 1 provides an unbiased estimate of the error rate (technically leave-one-out estimators are almost unbiased (Vapnik, 1998)).

3.2.2 Train-test setting: Given a training set of $X = (x_i)_{i=1}^n$ expression profiles and labels $(y_i)_{i=1}^n$, a test set of $Z = (x_j)_{j=1}^{n'}$ expression profiles with labels $(t_j)_{j=1}^{n'}$, and gene sets $\Gamma = \{\gamma_1, ..., \gamma_m\}$ the procedure outlined in Algorithm 2 provides an unbiased error estimate on the test set.

Algorithm 1: Leave-one-out procedure for pathway based classification.

input : training data and gene sets

return: error rate

for $i = 1$ **to** n **do**
    split the data into x_i (the i-th data point) and $X^{\setminus i}$ (the data with the i-th point removed);
    compute Tr $= ES_i^k$ for $X^{\setminus i}$ (this is the enrichment of the m gene sets on the $n - 1$ data in $X^{\setminus i}$);
    compute Test $= ES_i^k$ for x_i (this is the enrichment of the m gene sets on x_i, the label i-th point is not used in the estimation of the enrichment score by leaving this point out of the Parzen estimator);
    use Tr to build logistic regression with variable selection model M_i;
    apply M_i to Test to obtain prediction $\hat{y}_i$;
    if $y_i \neq \hat{y}_i$ **then** error rate $=$ error rate $+ 1$;

return error rate

Algorithm 2: Test error estimate for pathway based classification.

input : training data, test data, and gene sets

return: error rate

compute Tr $= ES_i^k$ for X (this is the enrichment of the m gene sets on the training data X);
use Tr to build logistic regression with variable selection model M;
for $j = 1$ **to** n' **do**
    compute Test $= ES_j^k$ for z_j (this is the enrichment of the m gene sets on the j-th test sample, use only the training data X to compute the Parzen estimator);
    apply M to Test to obtain prediction $\hat{t}_j$;
    if $t_j \neq \hat{t}_j$ **then** error rate $=$ error rate $+ 1$;
return error rate

3.3 Computation of Absorption Probabilities

To compute the correlation coefficients in the nonparametric model we need to compute the probability that the expression of the j-th gene in the i-th sample is representative of class 1 or class 2, $\mathbb{P}(x_j^i \in C_1)$ and $\mathbb{P}(x_j^i \in C_2)$ for all samples $i = 1, ..., n$ and genes $j = 1, .., p$. We first scale the expression data for each gene to $[0, 1]$, $\hat{x}_j^i = \frac{x_j^i - \min_i(x_j^i)}{\max_i(x_j^i) - \min_i(x_j^i)}$. The class membership probabilities are the probabilities of absorption to the left or right extreme, which are $\{0, 1\}$ for the scaled data, depending on whether $\hat{x}_j^i$ is greater or less than the scaled class means (see Table 4). This simply reflects the directionality assumption of our model.

Let

$$u(\hat{x}) = \mathbb{P}(\text{absorption at 0 starting at } \hat{x}),$$
$$v(\hat{x}) = \mathbb{P}(\text{absorption at 1 starting at } \hat{x}),$$

$\mathbb{P}(\hat{x}_j^i \in C_1)$	$\hat{x}_j^i \leq \hat{\mu}_{j1}$	$\mathbb{P}(\text{absorption at 0 starting at } \hat{x}_j^i \vert \hat{p}_{j1})$
	$\hat{x}_j^i > \hat{\mu}_{j1}$	$\mathbb{P}(\text{absorption at 1 starting at } \hat{x}_j^i \vert \hat{p}_{j1})$
$\mathbb{P}(\hat{x}_j^i \in C_2)$	$\hat{x}_j^i \leq \hat{\mu}_{j2}$	$\mathbb{P}(\text{absorption at 0 starting at } \hat{x}_j^i \vert \hat{p}_{j2})$
	$\hat{x}_j^i > \hat{\mu}_{j2}$	$\mathbb{P}(\text{absorption at 1 starting at } \hat{x}_j^i \vert \hat{p}_{j2})$

Table 4. Probability of class membership as a function of $\hat{x}_j^i$ and the class means.

and let $p(\hat{x})$ be supported on $[0, 1]$, then

$$\mathbb{P}(\text{absorption at 0 starting at } \hat{x} \vert p(\hat{x})) = \int_0^{\hat{x}} u(\hat{x}) \, p(\hat{x}) \, d\hat{x},$$

$$\mathbb{P}(\text{absorption at 1 starting at } \hat{x} \vert p(\hat{x})) = \int_{\hat{x}}^1 v(\hat{x}) \, p(\hat{x}) \, d\hat{x}.$$

The absorption probabilities of a Brownian motion at the end points of a line segment can be computed by solving the heat equation with appropriate boundary conditions, the Dirichlet problem (Durrett, 1996). In the above case we solve for

$$\frac{d^2 u(\hat{x})}{d\hat{x}^2} = 0 \quad s.t. \quad u(0) = 0, \; u(1) = 1$$

$$\frac{d^2 v(\hat{x})}{d\hat{x}^2} = 0 \quad s.t. \quad v(0) = 1, \; v(1) = 0.$$

This results in the solutions

$$u(\hat{x}) = \hat{x}, \qquad v(\hat{x}) = 1 - \hat{x}.$$

Given the Parzen estimates of the densities for the two classes

$$\hat{p}_{j1}(\hat{x}) = \frac{1}{n_1 \sigma_{j1} \sqrt{2\pi}} \sum_{i \in C_1} e^{-\vert \hat{x}_j^i - \hat{x} \vert^2 / 2\sigma_{j1}^2},$$

$$\hat{p}_{j2}(\hat{x}) = \frac{1}{n_2 \sigma_{j2} \sqrt{2\pi}} \sum_{i \in C_2} e^{-\vert \hat{x}_j^i - \hat{x} \vert^2 / 2\sigma_{j2}^2},$$

we can compute the absorption probabilities as

$$\mathbb{P}(\text{absorption at 0 starting at } \hat{x} \vert \hat{p}_{jc}) = \int_0^{\hat{x}} s \, \hat{p}_{jc}(s) \, ds$$

$$\mathbb{P}(\text{absorption at 1 starting at } \hat{x} \vert \hat{p}_{jc}) = \int_{\hat{x}}^1 (1 - s) \, \hat{p}_{jc}(s) \, ds$$

where c denotes the classes $\{1, 2\}$. Solving the integrals results in a weighted sum of error functions and exponentials.

4 DISCUSSION

In this paper we introduce a formal statistical method to measure the enrichment of each sample in an expression data set with respect to a priori defined gene sets. This allows us to assay the natural variation of pathway activity in observed gene expression data sets. It is a natural extension of methods that measure the enrichment of an entire data set with respect to a priori defined gene sets (Subramanian *et al.*, 2005; Barry *et al.*, 2005; Kim and Volsky, 2005; Tomfohr *et al.*, 2005).

The method was validated on a variety of model systems: oncogenic cell lines, mouse models, and known gender differences in expression. The utility of the method was demonstrated by clustering and building classification models in the space of pathways or gene sets. These were in general as accurate as methods applied in the space of genes but more interpretable and robust. This robustness was illustrated by the ability to build models between different expression based technologies— cross-platform models. This is hard to do in the single gene setting.

A variety of open questions regarding the pathway paradigm and our implementation of it remain. Some of these questions are technical and some are fundamental with respect to both statistical analysis and molecular biology.

We first discuss the technical issues:

- Enrichment statistic: We use a maximum deviation statistic to compute our enrichment score. The theory behind BLAST (Ewans and Grant, 2002) offers insights as to how we may improve our statistic by adding to the maximal extremal excursion the top r excursions. This would especially make sense when the gene set corresponds to genes in a pathway that subdivide into sub-pathways, some of which are up regulated and some of which are down regulated.

- Correlation statistic: We used a Brownian motion model to compute our correlation statistic. This outperformed a simple Gaussian model and a model based upon the cumulative distribution function of the Parzen estimator. However, these models are by no means exhaustive and other statistics may be as robust but with greater sensitivity.

- Extension to real-valued phenotypes: We stated the procedure for the case with binary phenotypes. The crux of an extension to real-valued phenotypes would be the computation of an appropriate correlation statistic. In the context of a survival model this would not be difficult but in general it can be complicated.

There are two fundamental questions with respect to our approach and they are intimately related

- What is a pathway (gene set): Gene sets can be derived from experimental perturbations, literature based studies, and a variety of other origins. A fundamental question is which of these sets is most appropriate. For example, a database of gene sets may contain 5 Ras pathways from a variety of experiments or literature surveys. For a particular analysis which is most appropriate? The authors believe that a partial answer or consensus is developing that experimentally based gene sets are in general more robust than ones derived from literature. However, the quantification of this and a statistically formal method for scoring gene sets is still an open problem.

- Likelihood based testing: The statistic used in our hypothesis testing framework is likelihood based, $\mathbb{P}(\text{data}\vert\text{pathway})$. The problem with using this likelihood based framework is that in this formulation is that the pathway we condition upon is not fixed. The Ras pathway as defined today is different than the Ras pathway as defined in two weeks, some genes are added and some removed. In the above framework one can then ask which pathway are we testing, is there multiplicity in the Ras pathway and if so how many Ras pathways are there. An alternative approach which is conceptually very appealing is to build our statistical framework on the posterior, $\mathbb{P}(\text{pathway}\vert\text{data})$. This provides a uniform framework and quantity that we can use to score the different Ras pathways in the previous example. The fundamental problem in using the posterior is that a prior is needed on the space of pathways (for example priors over possible Ras pathways). The construction or estimation by sampling gene expression data sets of such a priori defined gene sets starting with a database of pathways is

a very interesting and challenging computational biology and statistics problem.

ACKNOWLEDGEMENT

REFERENCES

Alvarez, J., Febbo, P., Ramaswamy, S., Loda, M., Richardson, A. and Frank, D. (2005) Identification of a genetic signature of activated signal transducer and activator of transcription 3 in human tumors. *Cancer Res*, **65**, 5054–62.

Barry, W., Nobel, A. and Wright, F. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–9.

Bild, A., Yao, G., Chang, J., Wang, Q., Potti, A., Chasse, D., Joshi, M., Harpole, D., Lancaster, J., Berchuck, A., Olson, J., Marks, J., Dressman, H., West, M. and Nevins, J. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Black, E., Huang, E., Dressman, H., Rempel, R., Laakso, N., Asa, S., Ishida, S., West, M. and Nevins, J. (2003) Distinct gene expression phenotypes of cells lacking Rb and Rb family members. *Cancer Res*, **63**, 3716–23.

Brunet, J., Tamayo, P., Golub, T. and Mesirov, J. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, **101**, 4164–9.

Dhanasekaran, S., Barrett, T., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K., Rubin, M. and Chinnaiyan, A. (2001) Delineation of prognostic biomakers in prostate cancer. *Nature*, **412**, 822–826.

Durrett, R. (1996) *Stochastic Calculus: A Practical Introduction*. CRC Press, Boca Raton, FL.

Ewans, W. and Grant, G. (2002) *Statistical Methods in Bioinformatics*. Springer.

Febbo, P., Richie, J., George, D., Loda, M., Manola, J., Shankar, S., Barnes, A., Tempany, C., Catalona, W., Kantoff, P. and Oh, W. (2005) Neoadjuvant docetaxel before radical prostatectomy in patients with high-risk localized prostate cancer. *Clin Cancer Res*, **11**, 5233–40.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications, Vol 1*. John Wiley & Sons, New York.

Garber, M., Troyanskaya, O., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G., Perou, C., Whyte, R., Altman, R., Brown, P., Botstein, D. and Petersen, I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A*, **98**, 13784–9.

Glinsky, G., Glinskii, A., Stephenson, A., Hoffman, R. and Gerald, W. (2004) Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest*, **113**, 913–23.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–7.

Hastie, T., Tibshirani, R. and Friedman, J. (2000) *The Elements of Statistical Learning, Data Mining, Inference and Prediction.* Springer Verlag, New York.

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., D'Amico, M., Pestell, R., West, M. and Nevins, J. (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet*, **34**, 226–30.

Kim, S. and Volsky, D. (2005) PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics*, **6**.

Majumder, P., Febbo, P., Bikoff, R., Berger, R., Xue, Q., McMahon, L., Manola, J., Brugarolas, J., McDonnell, T., Golub, T., Loda, M., Lane, H. and Sellers, W. (2004) mTOR inhibition reverses Akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways. *Nat Med*, **10**, 594–601.

Mootha, V., Lindgren, C., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M., Patterson, N., Mesirov, J., Golub, T., Tamayo, P., Spiegelman, B., Lander, E., Hirschhorn, J., Altshuler, D. and Groop, L. (2003) Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**, 267–73.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E. and Golub, T. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–42.

Potti, A., Mukherjee, S., Petersen, R., Dressman, H., Bild, A., Koontz, J., Kratzke, R., Watson, M., Kelley, M., Ginsburg, G., West, M., Haropole, D. J. and Nevins, J. (2006) A Genomic Strategy to Refine Prognosis in Early Stage Non-Small Cell Lung Carcinoma. Submitted.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, J. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J., Ladd-Acosta, C., Mesirov, J., Golub, T. and Jacks, T. (2005) An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat Genet*, **37**, 48–55.

Tomfohr, J., Lu, J. and Kepler, T. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**.

Vapnik, V. (1998) *Statistical learning theory*. J. Wiley and Sons.

Warnat, P., Eils, R., and Brors, B. (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, **6**.

Welsh, M., Sapinoso, L., Su, A., Kern, S., Wang-Rodriguez, J., Moskaluk, C., Frierson, H. J. and Hampton, G. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.*, **61**, 5974–5978.