Application of *a priori* established gene sets to discover biologically important differential expression in microarray data

Andrea Bild*† and Phillip George Febbo*†‡§

**Duke Institute for Genome Sciences and Policy and Departments of ‡Medicine and †Molecular Genetics and Microbiology, Duke University Medical Center, Duke University, Durham, NC 27710*

From inception, microarray analysis has facilitated discovery by associating gene expression with biological and/or clinical sample characteristics. However, gleaning biological insight from the long lists of genes generated by microarray analysis remains a significant challenge. In this issue of PNAS, Subramanian *et al.* (1) describe and validate gene set enrichment analysis (GSEA), a computational method that helps rapidly connect gene expression with biology and promises to be a valuable addition to publicly available computational resources.

Early on, investigators adapted unsupervised computational methods such as hierarchical clustering (2) and self-organized maps (3) to arrange genes and samples in groups or clusters based solely on the similarity of their gene expression. These methods successfully revealed the orchestrated gene expression underlying basic cellular processes such as yeast replication (4), fibroblast cell proliferation (5), and hematopoietic differentiation (3), and they continue to be used widely today. Unsupervised methods are unbiased and remain important tools for class discovery.

Alternatively, supervised methods of analysis use sample classifiers along with gene expression to rapidly identify hypothesis-driven correlations (i.e., tumor v. normal, pathological grade, recurrent disease, histological category, etc.). A few examples of supervised methods of analysis include significance analysis of microarray (SAM) (6), class prediction (7), support vector machines (8), and probit regression analysis (9, 10). In the field of oncology, supervised methods of gene expression have successfully identified novel marker genes for diagnosis (11), prognosis (12), and therapeutic response (13). Supervised methods can help overcome obfuscating technical or biological variation in gene expression and continue to identify important associations between sample phenotypes and gene expression.

GSEA represents an innovative method of supervised analysis. This analysis is performed by (*i*) ranking all genes in the data set based on their correlation to the chosen phenotype, (*ii*) identifying the rank positions of all members of the gene set, and (*iii*) calculating an enrichment score (ES) that represents the difference between the observed rankings and that which would be expected assuming a random rank distribution (see figure 1 A and B in ref. 1). After establishing the ES for each gene set across the phenotype, GSEA reiteratively randomizes the sample labels and retests for enrichment across the random classes. By performing repeated class label randomizations, the ES for each gene set across the true classes can be compared to the ES distribution from the random classes. Those gene sets that significantly outperform iterative random class permutations are considered significant.

Pathway-oriented approaches similar to GSEA have recently become more popular (reviewed in ref. 14). As a general approach, these methods use predetermined aggregations of genes (alternatively called gene sets, metagenes, or gene modules) rather than individual genes to assess for coordinate expression within samples or sample classifications. Investigators have pursued multiple strategies to develop informative gene sets; some groups organize gene sets based on public sources (i.e., KEGG pathways), and others perform experiments to define gene sets.

Successful gene sets can help identify underlying genetic abnormalities or signal transduction networks driving disease pathologies and help effectively bridge microarray data with biological significance. Published reports using aggregated gene sets have identified oxidative phosphorylation pathway deregulation in human diabetic muscle (15), Myc, Ras, and Rb pathway deregulation within murine tumor models (9, 10), Hif1a inhibition in a murine prostate cancer model treated with mTOR inhibition (16), and *k-RAS* deregulation in lung adenocarcinoma (17).

By capitalizing on the statistical advantage gained by established gene associations, relatively small individual differential gene expression can combine to create strong correlations between a gene set and a class distinction. This advantage is demonstrated by Subramanian *et al.* (1) when GSEA is applied to determine differential gene expression between lymphoblastoid cell lines derived from either men or women. Using gene sets aggregated based on cytogenic location, Subramanian *et al.* successfully identify differential expression of genes located on the Y chromosome. In addition, a gene set containing genes known to escape X inactivation is significantly enriched in the female cell lines compared with the male cell lines. Thus, even when the differential expression of individual genes is likely to be 2-fold or less (i.e., escape from X inactivation) and the genes of interest constitute a small minority of all genes assayed, GSEA can detect differential expression.

For heterogeneous samples and/or when there are relatively subtle differences between sample classes, standard supervised or unsupervised methods may fail to detect differential gene expression. By contributing additional information (i.e., associations between genes), GSEA provides investigators with a method that can reveal biologically meaningful differential expression when standard methods fail. In a previously published application of GSEA (15), the use of aggregated gene sets identified differential activity of a gene set for oxidative phosphorylation despite an average difference in gene expression of 20% for gene set members between diabetic and nondiabetic samples.

Subramanian *et al.* (1) identify differential activity of gene sets for cellular proliferation and amino acid biosynthesis between lung tumors with "good" or "poor" outcomes in two independent data sets. As in the diabetic study, when each gene's expression is treated independently during supervised analysis, no gene is significantly differentially expressed between good and poor outcome lung samples. In addition, there is remarkably little overlap between the top ranked genes for the two data sets (only 12 genes of the top 100 genes with expression correlating to poor outcome). This lack of overlap, frequently observed between independent cancer data sets, is likely due to the cumulative effects of each group's approach to sample collection, tumor dissection, RNA

Conflict of interest statement: No conflicts declared.

See companion article on page 15545.

§To whom correspondence should be addressed. E-mail: phil.febbo@duke.edu.

isolation, target preparation, microarray hybridization, and microarray scanning, as well as sampling differences between two moderately sized sets of tumors.

GSEA was able to overcome these differences and identify gene sets with significant differential expression in both data sets. Although the pathways identified remain to be validated, this observation suggests that GSEA is less sensitive to unavoidable technical and sampling differences between independently collected samples. This finding raises the additional possibility that computational approaches such as GSEA that use aggregated gene sets may be better suited to apply preliminary findings across technical platforms, for example, taking findings based upon frozen tissue and applying them to findings from paraffin-embedded tissues (one of the major current challenges for microarray studies).

Additional examples from the current article further underscore the potential of GSEA to interrogate and deconvolute the disruption of multiple basic cellular signaling pathways occurring during oncogenesis. The authors applied GSEA to a previously published data set of acute lymphoid leukemia (ALL, $n = 24$) and acute myeloid leukemia (AML, $n = 24$) using their cytogenetic gene sets and identified five cytogenetic abnormalities (5q31, 17q23, 13q14, 6q12, and 14q32). With relatively few clinical samples, they were able to highlight fundamental genetic alterations previously described in hematopoietic malignancies. The authors also used expression data from the NCI-60 set of cancer cell lines, GSEA, and their "leading edge" analysis to determine that the mitogen-activated protein-kinase signaling pathway is differentially expressed between p53 positive and negative cell lines. In "leading edge" analysis, the genes shared across the gene sets most strongly associated with the phenotype are identified. Because signaling pathways can have significant overlap with respect to effecter proteins, "leading edge" analysis has the potential to identify the most critical subpathway. In aggregate, the associations between gene sets and clinical or molecular phenotypes demonstrated in Subramanian *et al.* (1) underscore the ability of GSEA to connect gene expression data with biological insight.

GSEA has evolved from the initial application in Mootha *et al.* (15) to the current detailed reporting of the methodology. First, the current iteration of GSEA normalizes the ES based on the number of gene members in the data set; this partially addresses prior criticism (18) and more readily allows comparison across gene sets of different sizes. Next, whereas GSEA initially put equal statistical weight on each step (i.e., the location of each gene set member in the ranking of all genes according to phenotype), greater weight is now placed on genes with stronger correlation with the phenotype (i.e., genes located near the top or bottom of the rankings). Finally, they have revised the measure of significance and use a calculated false discovery rate (FDR) that is based upon the distribution of results during repetitive, random assignments of class designations. Subramanian *et al.* (1) report the FDR to be less conservative than their preliminary implementation using family-wise error rate, a change justified by GSEA's role as a hypothesis-generating analytic approach. Importantly, despite these changes, the current methods of GSEA reproduce the earlier findings by Mootha *et al.* (15).

As is ideal for novel computational methodologies with high potential impact on microarray analysis, Subramanian *et al.* (1) have created a JAVA-based program executable using Windows, Macintosh, or Unix/Linux called GSEA-P. Access to the software is free, and Subramanian *et al.* also provide an inventory of gene sets called MSigDB. As individual investigators adopt GSEA, they will be able to immediately interrogate established data sets using MSigDB. As independent groups adopt GSEA, test self-generated gene sets, and ideally contribute informative gene sets, MSigDB is likely to evolve into a valuable, living resource.

GSEA will continue to evolve as it becomes more widely adopted in the scientific community. GSEA is absolutely dependent on the quality of gene sets and determining the genes that best represent a pathway's activity is complex. Subramanian *et al.* (1) have developed gene sets opportunistically, and the most valid methods for defining gene sets is likely to be an area of continued investigation. In addition, although the specific statistical weights and measured used by Subramanian *et al.* to assess the correlation between the expression of a set of genes and a sample phenotype represent a valuable starting point, there will undoubtedly be additional computational exploration and evaluation. Finally, if GSEA is going to be used to assess pathway activity in samples so that treatment can be targeted (a logical and potentially powerful application of this methodology), the methods have to be refined to allow the scoring of individual samples.

GSEA uses aggregated gene sets to identify biological processes present across phenotypes in microarray data sets and represents the latest approach to gene expression analysis. These modular approaches facilitate a greater understanding of the underlying biology driving pathological phenotypes, and they promise to facilitate significant contributions toward the molecular characterization of human disease.

We thank Joseph Nevins, Bala Balakumarin, Geoffrey Ginsburg, and Alessandro Porrello for their careful review of this Commentary. P.G.F. is a Damon Runyon Cancer Research Foundation Clinical Investigator and receives additional support from National Institutes of Health Grant CA-089031 and the Prostate Cancer Foundation.

1. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 15545–15550.
2. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
3. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. & Golub, T. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2907–2912.
4. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998) *Mol. Biol. Cell* **9,** 3273–3297.
5. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., *et al.* (1999) *Science* **283,** 83–87.
6. Tusher, V. G., Tibshirani, R. & Chu, G. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 5116–5121.
7. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.* (1999) *Science* **286,** 531–537.
8. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. & Haussler, D. (2000) *Bioinformatics* **16,** 906–914.
9. Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., D'Amico, M., Pestell, R. G., West, M. & Nevins, J. R. (2003) *Nat. Genet.* **34,** 226–230.
10. Black, E. P., Huang, E., Dressman, H., Rempel, R., Laakso, N., Asa, S. L., Ishida, S., West, M. & Nevins, J. R. (2003) *Cancer Res.* **63,** 3716–3723.
11. Rubin, M. A., Zhou, M., Dhanasekaran, S. M., Varambally, S., Barrette, T. R., Sanda, M. G., Pienta, K. J., Ghosh, D. & Chinnaiyan, A. M. (2002) *J. Am. Med. Assoc.* **287,** 1662–1670.
12. van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., *et al.* (2002) *N. Engl. J. Med.* **347,** 1999–2009.
13. Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A. & Staudt, L. M. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9991–9996.
14. Segal, E., Yelensky, R. & Koller, D. (2003) *Bioinformatics* **19,** Suppl. 1, i273–i282.
15. Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al.* (2003) *Nat. Genet.* **34,** 267–273.
16. Majumder, P. K., Febbo, P. G., Bikoff, R., Berger, R., Xue, Q., McMahon, L. M., Manola, J., Brugarolas, J., McDonnell, T. J., Golub, T. R., *et al.* (2004) *Nat. Med.* **10,** 594–601.
17. Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J. J., Ladd-Acosta, C., Mesirov, J., Golub, T. R. & Jacks, T. (2005) *Nat. Genet.* **37,** 48–55.
18. Attie, A. D. & Kendziorski, C. M. (2003) *Nat. Genet.* **34,** 244–245.