Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

[a]Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141; [c]Department of Systems Biology, Alpert 536, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02446; [d]Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine, and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708; [e]Department of Medical Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [f]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; [g]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C2-023, P.O. Box 19024, Seattle, WA 98109-1024; [h]Department of Neurology, Enders 260, Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115; [i]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; and [j]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA 02142

Contributed by Eric S. Lander, August 2, 2005

Although genomewide RNA expression analysis has become a routine tool in biomedical research, extracting biological insight from such information remains a major challenge. Here, we describe a powerful analytical method called Gene Set Enrichment Analysis (GSEA) for interpreting gene expression data. The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation. We demonstrate how GSEA yields insights into several cancer-related data sets, including leukemia and lung cancer. Notably, where single-gene analysis finds little similarity between two independent studies of patient survival in lung cancer, GSEA reveals many biological pathways in common. The GSEA method is embodied in a freely available software package, together with an initial database of 1,325 biologically defined gene sets.

microarray

Genomewide expression analysis with DNA microarrays has become a mainstay of genomics research (1, 2). The challenge no longer lies in obtaining gene expression profiles, but rather in interpreting the results to gain insights into biological mechanisms.

In a typical experiment, mRNA expression profiles are generated for thousands of genes from a collection of samples belonging to one of two classes, for example, tumors that are sensitive vs. resistant to a drug. The genes can be ordered in a ranked list L, according to their differential expression between the classes. The challenge is to extract meaning from this list.

A common approach involves focusing on a handful of genes at the top and bottom of L (i.e., those showing the largest difference) to discern telltale biological clues. This approach has a few major limitations.

(*i*) After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are modest relative to the noise inherent to the microarray technology.

(*ii*) Alternatively, one may be left with a long list of statistically significant genes without any unifying biological theme. Interpretation can be daunting and ad hoc, being dependent on a biologist's area of expertise.

(*iii*) Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting in concert. An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.

(*iv*) When different groups study the same biological system, the list of statistically significant genes from the two studies may show distressingly little overlap (3).

To overcome these analytical challenges, we recently developed a method called Gene Set Enrichment Analysis (GSEA) that evaluates microarray data at the level of gene sets. The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways or coexpression in previous experiments. The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction.

We used a preliminary version of GSEA to analyze data from muscle biopsies from diabetics vs. healthy controls (4). The method revealed that genes involved in oxidative phosphorylation show reduced expression in diabetics, although the average decrease per gene is only 20%. The results from this study have been independently validated by other microarray studies (5) and by *in vivo* functional studies (6).

Given this success, we have developed GSEA into a robust technique for analyzing molecular profiling data. We studied its characteristics and performance and substantially revised and generalized the original method for broader applicability.

In this paper, we provide a full mathematical description of the GSEA methodology and illustrate its utility by applying it to several diverse biological problems. We have also created a software package, called GSEA-P and an initial inventory of gene sets (Molecular Signature Database, MSigDB), both of which are freely available.

Methods

Overview of GSEA. GSEA considers experiments with genomewide expression profiles from samples belonging to two classes, labeled 1 or 2. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric (Fig. 1*A*).

Given an *a priori* defined set of genes S (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom. We expect

Freely available online through the PNAS open access option.

Abbreviations: ALL, acute lymphoid leukemia; AML, acute myeloid leukemia; *ES*, enrichment score; FDR, false discovery rate; GSEA, Gene Set Enrichment Analysis; MAPK, mitogen-activated protein kinase; MSigDB, Molecular Signature Database; NES, normalized enrichment score.

See Commentary on page 15278.

[b]A.S. and P.T. contributed equally to this work.

[k]To whom correspondence may be addressed. E-mail: lander@broad.mit.edu or mesirov@broad.mit.edu.

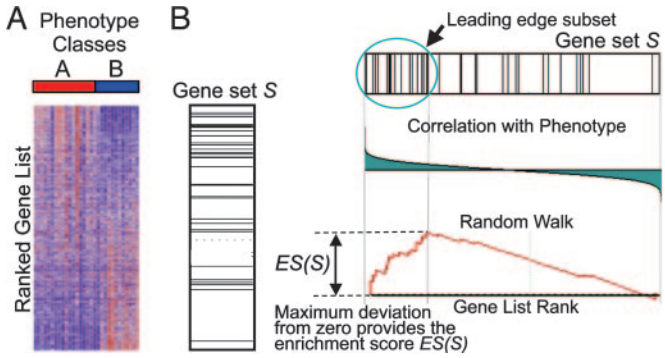

Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the ''gene tags,'' i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

that sets related to the phenotypic distinction will tend to show the latter distribution.

There are three key elements of the GSEA method:

Step 1: Calculation of an Enrichment Score. We 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*. The score is calculated by walking down the list *L*, increasing a running-sum statistic when we encounter a gene in *S* and decreasing it when we encounter genes not in *S*. The magnitude of the increment depends on the correlation of the gene with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov–Smirnov-like statistic (ref. 7 and Fig. 1*B*).

Step 2: Estimation of Significance Level of *ES*. We estimate the statistical significance (nominal *P* value) of the *ES* by using an empirical phenotype-based permutation test procedure that preserves the complex correlation structure of the gene expression data. Specifically, we permute the phenotype labels and recompute the *ES* of the gene set for the permuted data, which generates a null distribution for the *ES*. The empirical, nominal *P* value of the observed *ES* is then calculated relative to this null distribution. Importantly, the permutation of class labels preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes.

Step 3: Adjustment for Multiple Hypothesis Testing. When an entire database of gene sets is evaluated, we adjust the estimated significance level to account for multiple hypothesis testing. We first normalize the *ES* for each gene set to account for the size of the set, yielding a normalized enrichment score (*NES*). We then control the proportion of false positives by calculating the false discovery rate (FDR) (8, 9) corresponding to each *NES*. The FDR is the estimated probability that a set with a given *NES* represents a false positive finding; it is computed by comparing the tails of the observed and null distributions for the *NES*.

Table 1. *P* value comparison of gene sets by using original and new methods

Gene set	Original method nominal *P* value	New method nominal *P* value
S1: chrX inactive	0.007	<0.001
S2: vitcb pathway	0.51	0.38
S3: nkt pathway	0.023	0.54

The details of the implementation are described in the *Appendix* (see also *Supporting Text*, which is published as supporting information on the PNAS web site).

We note that the GSEA method differs in several important ways from the preliminary version (see *Supporting Text*). In the original implementation, the running-sum statistic used equal weights at every step, which yielded high scores for sets clustered near the middle of the ranked list (Fig. 2 and Table 1). These sets do not represent biologically relevant correlation with the phenotype. We addressed this issue by weighting the steps according to each gene's correlation with a phenotype. We noticed that the use of weighted steps could cause the distribution of observed *ES* scores to be asymmetric in cases where many more genes are correlated with one of the two phenotypes. We therefore estimate the significance levels by considering separately the positively and negatively scoring gene sets (*Appendix*; see also Fig. 4, which is published as supporting information on the PNAS web site).

Our preliminary implementation used a different approach, familywise-error rate (FWER), to correct for multiple hypotheses testing. The FWER is a conservative correction that seeks to ensure that the list of reported results does not include even a single false-positive gene set. This criterion turned out to be so conservative that many applications yielded no statistically significant results. Because our primary goal is to generate hypotheses, we chose to use the FDR to focus on controlling the probability that each reported result is a false positive.

Based on our statistical analysis and empirical evaluation, GSEA shows broad applicability. It can detect subtle enrichment signals and it preserves our original results in ref. 4, with the oxidative phosphorylation pathway significantly enriched in the normal samples ($P = 0.008$, $FDR = 0.04$). This methodology has been implemented in a software tool called GSEA-P.

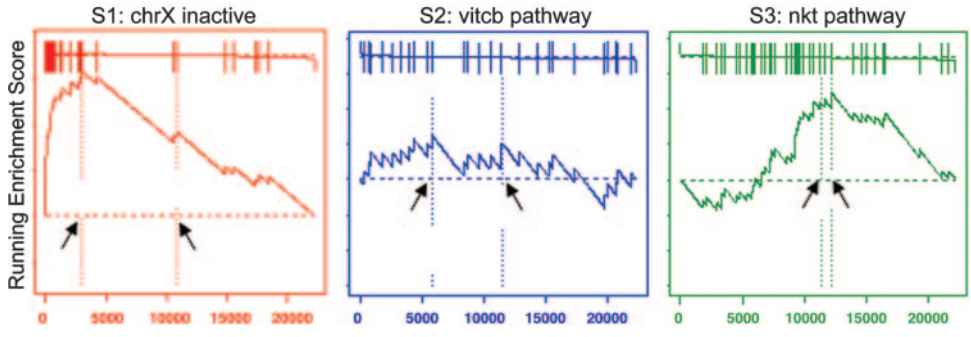

Fig. 2. Original (4) enrichment score behavior. The distribution of three gene sets, from the C2 functional collection, in the list of genes in the male/female lymphoblastoid cell line example ranked by their correlation with gender: S1, a set of chromosome X inactivation genes; S2, a pathway describing vitamin c import into neurons; S3, related to chemokine receptors expressed by T helper cells. Shown are plots of the running sum for the three gene sets: S1 is significantly enriched in females as expected, S2 is randomly distributed and scores poorly, and S3 is not enriched at the top of the list but is nonrandom, so it scores well. Arrows show the location of the maximum enrichment score and the point where the correlation (signal-to-noise ratio) crosses zero. Table 1 compares the nominal *P* values for S1, S2, and S3 by using the original and new method. The new method reduces the significance of sets like S3.

The Leading-Edge Subset. Gene sets can be defined by using a variety of methods, but not all of the members of a gene set will typically participate in a biological process. Often it is useful to extract the core members of high scoring gene sets that contribute to the *ES*. We define the leading-edge subset to be those genes in the gene set *S* that appear in the ranked list *L* at, or before, the point where the running sum reaches its maximum deviation from zero (Fig. 1*B*). The leading-edge subset can be interpreted as the core of a gene set that accounts for the enrichment signal.

Examination of the leading-edge subset can reveal a biologically important subset within a gene set as we show below in our analysis of P53 status in cancer cell lines. This approach is especially useful with manually curated gene sets, which may represent an amalgamation of interacting processes. We first observed this effect in our previous study (4) where we manually identified two high scoring sets, a curated pathway and a computationally derived cluster, which shared a large subset of genes later confirmed to be a key regulon altered in human diabetes.

High scoring gene sets can be grouped on the basis of leading-edge subsets of genes that they share. Such groupings can reveal which of those gene sets correspond to the same biological processes and which represent distinct processes.

The GSEA-P software package includes tools for examining and clustering leading-edge subsets (*Supporting Text*).

Variations of the GSEA Method. We focus above and in *Results* on the use of GSEA to analyze a ranked gene list reflecting differential expression between two classes, each represented by a large number of samples. However, the method can be applied to ranked gene lists arising in other settings.

Genes may be ranked based on the differences seen in a small data set, with too few samples to allow rigorous evaluation of significance levels by permuting the class labels. In these cases, a *P* value can be estimated by permuting the genes, with the result that genes are randomly assigned to the sets while maintaining their size. This approach is not strictly accurate: because it ignores gene-gene correlations, it will overestimate the significance levels and may lead to false positives. Nonetheless, it can be useful for hypothesis generation. The GSEA-P software supports this option.

Genes may also be ranked based on how well their expression correlates with a given target pattern (such as the expression pattern of a particular gene). In Lamb *et al.* (10), a GSEA-like procedure was used to demonstrate the enrichment of a set of targets of cyclin D1 list ranked by correlation with the profile of cyclin D1 in a compendium of tumor types. Again, approximate *P* values can be estimated by permutation of genes.

An Initial Catalog of Human Gene Sets. GSEA evaluates a query microarray data set by using a collection of gene sets. We therefore created an initial catalog of 1,325 gene sets, which we call MSigDB 1.0 (*Supporting Text*; see also Table 3, which is published as supporting information on the PNAS web site), consisting of four types of sets.

Cytogenetic sets (C_1, 319 gene sets). This catalog includes 24 sets, one for each of the 24 human chromosomes, and 295 sets corresponding to cytogenetic bands. These sets are helpful in identifying effects related to chromosomal deletions or amplifications, dosage compensation, epigenetic silencing, and other regional effects.

Functional sets (C_2, 522 gene sets). This catalog includes 472 sets containing genes whose products are involved in specific metabolic and signaling pathways, as reported in eight publicly available, manually curated databases, and 50 sets containing genes coregulated in response to genetic and chemical perturbations, as reported in various experimental papers.

Regulatory-motif sets (C_3, 57 gene sets). This catalog is based on our recent work reporting 57 commonly conserved regulatory motifs in the promoter regions of human genes (11) and makes it possible to link changes in a microarray experiment to a conserved, putative cis-regulatory element.

Neighborhood sets (C_4, 427 gene sets). These sets are defined by expression neighborhoods centered on cancer-related genes.

This database provides an initial collection of gene sets for use with GSEA and illustrates the types of gene sets that can be defined, including those based on prior knowledge or derived computationally.

GSEA-P Software and MSigDB Gene Sets. To facilitate the use of GSEA, we have developed resources that are freely available from the Broad Institute upon request. These resources include the GSEA-P software, MSigDB 1.0, and accompanying documentation.

The software is available as (*i*) a platform-independent desktop application with a graphical user interface; (*ii*) programs in R and in JAVA that advanced users may incorporate into their own analyses or software environments; (*iii*) an analytic module in our GENEPATTERN microarray analysis package (available upon request) (*iv*) a future web-based GSEA server to allow users to run their own analysis directly on the web site. A detailed example of the output format of GSEA is available on the site, as well as in *Supporting Text*.

Results

We explored the ability of GSEA to provide biologically meaningful insights in six examples for which considerable background information is available. In each case, we searched for significantly associated gene sets from one or both of the subcatalogs C1 and C2 (see above). Table 2 lists all gene sets with an FDR ≤ 0.25.

Male vs. Female Lymphoblastoid Cells. As a simple test, we generated mRNA expression profiles from lymphoblastoid cell lines derived from 15 males and 17 females (unpublished data) and sought to identify gene sets correlated with the distinctions "male>female" and "female>male."

We first tested enrichment of cytogenetic gene sets (C_1). For the male>female comparison, we would expect to find the gene sets on chromosome Y. Indeed, GSEA produced chromosome Y and the two Y bands with at least 15 genes (Yp11 and Yq11). For the female>male comparison, we would not expect to see enrichment for bands on chromosome X because most X linked genes are subject to dosage compensation and, thus, not more highly expressed in females (12).

We next considered enrichment of functional gene sets (C_2). The analysis yielded three biologically informative sets. One consists of genes escaping X inactivation [merged from two sources (13, 14) that largely overlap], discovering the expected enrichment in female cells. Two additional sets consist of genes enriched in reproductive tissues (testis and uterus), which is notable inasmuch as mRNA expression was measured in lymphoblastoid cells. This result is not simply due to differential expression of genes on chromosomes X and Y but remains significant when restricted to the autosomal genes within the sets (Table 5, which is published as supporting information on the PNAS web site).

p53 Status in Cancer Cell Lines. We next examined gene expression patterns from the NCI-60 collection of cancer cell lines. We sought to use these data to identify targets of the transcription factor p53, which regulates gene expression in response to various signals of cellular stress. The mutational status of the p53 gene has been reported for 50 of the NCI-60 cell lines, with 17 being classified as normal and 33 as carrying mutations in the gene (15).

We first applied GSEA to identify functional gene sets (C_2) correlated with p53 status. The $p53^+ > p53^-$ analysis identified five sets whose expression is correlated with normal p53 function (Table 2). All are clearly related to p53 function. The sets are (*i*) a biologically annotated collection of genes encoding proteins in the p53-signaling pathway that causes cell-cycle arrest in response to DNA damage; (*ii*) a collection of downstream targets of p53 defined

Table 2. Summary of GSEA results with FDR ≤ 0.25

Gene set	FDR
Data set: Lymphoblast cell lines	
Enriched in males	
chrY	<0.001
chrYp11	<0.001
chrYq11	<0.001
Testis expressed genes	0.012
Enriched in females	
X inactivation genes	<0.001
Female reproductive tissue expressed genes	0.045
Data set: p53 status in NCI-60 cell lines	
Enriched in p53 mutant	
Ras signaling pathway	0.171
Enriched in p53 wild type	
Hypoxia and p53 in the cardiovascular system	<0.001
Stress induction of HSP regulation	<0.001
p53 signaling pathway	<0.001
p53 up-regulated genes	0.013
Radiation sensitivity genes	0.078
Data set: Acute leukemias	
Enriched in ALL	
chr6q21	0.011
chr5q31	0.046
chr13q14	0.057
chr14q32	0.082
chr17q23	0.071
Data set: Lung cancer outcome, Boston study	
Enriched in poor outcome	
Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216
Data set: Lung cancer outcome, Michigan study	
Enriched in poor outcome	
Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

For detailed results, see Table 4, which is published as supporting information on the PNAS web site.

by experimental induction of a temperature-sensitive allele of p53 in a lung cancer cell line; (*iii*) an annotated collection of genes induced by radiation, whose response is known to involve p53; (*iv*) an annotated collection of genes induced by hypoxia, which is known to act through a p53-mediated pathway distinct from the response pathway to DNA damage; and (*v*) an annotated collection of genes encoding heat shock-protein signaling pathways that protect cells from death in response to various cellular stresses.

The complementary analysis ($p53^- > p53^+$) identifies one significant gene set: genes involved in the Ras signaling pathway. Interestingly, two additional sets that fall just short of the significance threshold contain genes involved in the Ngf and Igf1 signaling pathways. To explore whether these three sets reflect a common biological function, we examined the leading-edge subset for each gene set (defined above). The leading-edge subsets consist of 16, 11, and 13 genes, respectively, with each containing four genes encoding products involved in the mitogen-activated protein kinase

Fig. 3. Leading edge overlap for p53 study. This plot shows the *ras*, *ngf*, and *igf1* gene sets correlated with $P53^-$ clustered by their leading-edge subsets indicated in dark blue. A common subgroup of genes, apparent as a dark vertical stripe, consists of MAP2K1, PIK3CA, ELK1, and RAF1 and represents a subsection of the MAPK pathway.

(MAPK) signaling subpathway (MAP2K1, RAF1, ELK1, and PIK3CA) (Fig. 3). This shared subset in the GSEA signal of the Ras, Ngf, and Igf1 signaling pathways points to up-regulation of this component of the MAPK pathway as a key distinction between the $p53^-$ and $p53^+$ tumors. (We note that a full MAPK pathway appears as the ninth set on the list.)

Acute Leukemias. We next sought to study acute lymphoid leukemia (ALL) and acute myeloid leukemia (AML) by comparing gene expression profiles that we had previously obtained from 24 ALL patients and 24 AML patients (16).

We applied GSEA to the cytogenetic gene sets (C_1), expecting that chromosomal bands showing enrichment in one class would likely represent regions of frequent cytogenetic alteration in one of the two leukemias. The ALL>AML comparison yielded five gene sets (Table 2), which could represent frequent amplification in ALL or deletion in AML. Indeed, all five regions are readily interpreted in terms of the current knowledge of leukemia.

The 5q31 band is consistent with the known cytogenetics of AML. Chromosome 5q deletions are present in most AML patients, with the critical region having been localized to 5q31 (17). The 17q23 band is a site of known genetic rearrangements in myeloid malignancies (18). The 13q14 band, containing the RB locus, is frequently deleted in AML but rarely in ALL (19). Finally, the 6q21 band contains a site of common chromosomal fragility and is commonly deleted in hematologic malignancies (20).

Interestingly, the remaining high scoring band is 14q32. This band contains the Ig heavy chain locus, which includes >100 genes expressed almost exclusively in the lymphoid lineage. The enrichment of 14q32 in ALL thus reflects tissue-specific expression in the lineage rather than a chromosomal abnormality.

The reciprocal analysis (AML>ALL) yielded no significantly enriched bands, which likely reflects the relative infrequency of deletions in ALL (21). The analyses with the cytogenetic gene sets thus show that GSEA is able to identify chromosomal aberrations common in particular cancer subtypes.

Comparing Two Studies of Lung Cancer. A goal of GSEA is to provide a more robust way to compare independently derived gene expression data sets (possibly obtained with different platforms) and obtain more consistent results than single gene analysis. To test robustness, we reanalyzed data from two recent studies of lung cancer reported by our own group in Boston (22) and another group in Michigan (23). Our goal was not to evaluate the results reported by the individual studies, but rather to examine whether common features between the data sets can be more effectively revealed by gene-set analysis rather than single-gene analysis.

Both studies determined gene-expression profiles in tumor samples from patients with lung adenocarcinomas ($n = 62$ for Boston; $n = 86$ for Michigan) and provided clinical outcomes (classified here as "good" or "poor" outcome). We found that no genes in either study were strongly associated with outcome at a significance level of 5% after correcting for multiple hypotheses testing.

From the perspective of individual genes, the data from the two studies show little in common. A traditional approach is to compare

the genes most highly correlated with a phenotype. We defined the gene set S_{Boston} to be the top 100 genes correlated with poor outcome in the Boston study and similarly $S_{Michigan}$ from the Michigan study. The overlap is distressingly small (12 genes in common) and is barely statistically significant with a permutation test ($P = 0.012$). When we added a Stanford study (24) involving 24 adenocarcinomas, the three data sets share only one gene in common among the top 100 genes correlated with poor outcome (Fig. 5 and Table 6, which are published as supporting information on the PNAS web site). Moreover, no clear common themes emerge from the genes in the overlaps to provide biological insight.

We then explored whether GSEA would reveal greater similarity between the Boston and Michigan lung cancer data sets. We compared the gene set from one data set, S_{Boston}, to the entire ranked gene list from the other. The set S_{Boston} shows a strong significant enrichment in the Michigan data ($NES = 1.90$, $P < 0.001$). Conversely, the poor outcome set $S_{Michigan}$ is enriched in the Boston data ($NES = 2.13$, $P < 0.001$). GSEA is thus able to detect a strong common signal in the poor outcome data (Fig. 6, which is published as supporting information on the PNAS web site).

Having found that GSEA is able to detect similarities between independently derived data sets, we then went on to see whether GSEA could provide biological insight by identifying important functional sets correlated with poor outcome in lung cancer. For this purpose, we performed GSEA on the Boston and Michigan data with the C_2 catalog of functional gene sets. Given the relatively weak signals found by conventional single-gene analysis in each study, it was not clear whether any significant gene sets would be found by GSEA. Nonetheless, we identified a number of genes sets significantly correlated with poor outcome (FDR $\leq$ 0.25): 8 in the Boston data and 11 in the Michigan data (Table 2). (The Stanford data had no genes or gene sets significantly correlated with outcome, which is most likely due to the smaller number of samples and many missing values in the data.)

Moreover, there is a large overlap among the significantly enriched gene sets in the two studies. Approximately half of the significant gene sets were shared between the two studies and an additional few, although not identical, were clearly related to the same biological process. Specifically, we found a set up-regulated by telomerase (25), two different tRNA synthesis-related sets, two different insulin-related sets, and two different p53-related sets. Thus, a total of 5 of 8 of the significant sets in Boston are identical or related to 6 of 11 in Michigan.

To provide greater insight, we next extended the analysis to include sets beyond those that met the FDR $\leq$ 0.25 criterion. Specifically, we considered the top scoring 20 gene sets in each of the three studies (60 gene sets) and their corresponding leading-edge subsets to better understand the underlying biology in the poor outcome samples (Table 4). Already in the Boston/Michigan overlap, we saw evidence of telomerase and p-53 response as noted above. Telomerase activation is believed to be a key aspect of pathogenesis in lung adenocarcinoma and is well documented as prognostic of poor outcome in lung cancer.

In all three studies, two additional themes emerge around rapid cellular proliferation and amino acid biosynthesis (Table 7, which is published as supporting information on the PNAS web site):

(*i*) We see striking evidence in all three studies of the effects of rapid cell proliferation, including sets related to Ras activation and the cell cycle as well as responses to hypoxia including angiogenesis, glycolysis, and carbohydrate metabolism. More than one-third of the gene sets (23 of 60) are related to such processes. These responses have been observed in malignant tumor microenvironments where enhanced proliferation of tumor cells leads to low oxygen and glucose levels (26). The leading-edge subsets of the associated significant gene sets include hypoxia-response genes such as HIF1A, VEGF, CRK, PXN, EIF2B1, EIF2B2, EIF2S2, FADD, NFKB1, RELA, GADD45A, and also Ras/MAPK activation genes (HRAS, RAF1, and MAP2K1).

(*ii*) We find strong evidence for the simultaneous presence of increased amino acid biosynthesis, *mTor* signaling, and up-regulation of a set of genes down-regulated by both amino acid deprivation and rapamycin treatment (27). Supporting this finding are 17 gene sets associated with amino acid and nucleotide metabolism, immune modulation, and *mTor* signaling. Based on these results, one might speculate that rapamycin treatment might have an effect on this specific component of the poor outcome signal. We note there is evidence of the efficacy of rapamycin in inhibiting growth and metastatic progression of non-small cell lung cancer in mice and human cell lines (28).

Our analysis shows that we find much greater consistency across the three lung data sets by using GSEA than by single-gene analysis. Moreover, we are better able to generate compelling hypotheses for further exploration. In particular, 40 of the 60 top scoring gene sets across these three studies give a consistent picture of underlying biological processes in poor outcome cases.

Discussion

Traditional strategies for gene expression analysis have focused on identifying individual genes that exhibit differences between two states of interest. Although useful, they fail to detect biological processes, such as metabolic pathways, transcriptional programs, and stress responses, that are distributed across an entire network of genes and subtle at the level of individual genes.

We previously introduced GSEA to analyze such data at the level of gene sets. The method was initially used to discover metabolic pathways altered in human diabetes and was subsequently applied to discover processes involved in diffuse large B cell lymphoma (29), nutrient-sensing pathways involved in prostate cancer (30), and in comparing the expression profiles of mouse to those of humans (31). In the current paper, we have refined the original approach into a sensitive, robust analytical method and tool with much broader applicability along with a large database of gene sets. GSEA can clearly be applied to other data sets such as serum proteomics data, genotyping information, or metabolite profiles.

GSEA features a number of advantages when compared with single-gene methods. First, it eases the interpretation of a large-scale experiment by identifying pathways and processes. Rather than focus on high scoring genes (which can be poorly annotated and may not be reproducible), researchers can focus on gene sets, which tend to be more reproducible and more interpretable. Second, when the members of a gene set exhibit strong cross-correlation, GSEA can boost the signal-to-noise ratio and make it possible to detect modest changes in individual genes. Third, the leading-edge analysis can help define gene subsets to elucidate the results.

Several other tools have recently been developed to analyze gene expression by using pathway or ontology information, e.g., (32–34). Most determine whether a group of differentially expressed genes is enriched for a pathway or ontology term by using overlap statistics such as the cumulative hypergeometric distribution. We note that this approach is not able to detect the oxidative phosphorylation results discussed above ($P = 0.08$, $FDR = 0.50$). GSEA differs in two important regards. First, GSEA considers all of the genes in an experiment, not only those above an arbitrary cutoff in terms of fold-change or significance. Second, GSEA assesses the significance by permuting the class labels, which preserves gene-gene correlations and, thus, provides a more accurate null model.

The real power of GSEA, however, lies in its flexibility. We have created an initial molecular signature database consisting of 1,325 gene sets, including ones based on biological pathways, chromosomal location, upstream cis motifs, responses to a drug treatment, or expression profiles in previously generated microarray data sets. Further sets can be created through genetic and chemical perturbation, computational analysis of genomic information, and additional biological annotation. In addition, GSEA itself could be used to refine manually curated pathways and sets by identifying the

leading-edge sets that are shared across diverse experimental data sets. As such sets are added, tools such as GSEA will help link prior knowledge to newly generated data and thereby help uncover the collective behavior of genes in states of health and disease.

Appendix: Mathematical Description of Methods

Inputs to GSEA.

1. Expression data set D with N genes and k samples.
2. Ranking procedure to produce Gene List L. Includes a correlation (or other ranking metric) and a phenotype or profile of interest C. We use only one probe per gene to prevent overestimation of the enrichment statistic (*Supporting Text*; see also Table 8, which is published as supporting information on the PNAS web site).
3. An exponent p to control the weight of the step.
4. Independently derived Gene Set S of N_H genes (e.g., a pathway, a cytogenetic band, or a GO category). In the analyses above, we used only gene sets with at least 15 members to focus on robust signals (78% of MSigDB) (Table 3).

Enrichment Score *ES*(*S*).

1. Rank order the N genes in D to form $L = \{g_1, \ldots, g_N\}$ according to the correlation, $r(g_j) = r_j$, of their expression profiles with C.
2. Evaluate the fraction of genes in S ("hits") weighted by their correlation and the fraction of genes not in S ("misses") present up to a given position i in L.

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$
$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}. \qquad [1]$$

The ES is the maximum deviation from zero of $P_{\text{hit}} - P_{\text{miss}}$. For a randomly distributed S, $ES(S)$ will be relatively small, but if it is concentrated at the top or bottom of the list, or otherwise nonrandomly distributed, then $ES(S)$ will be correspondingly high. When $p = 0$, *ES(S)* reduces to the standard Kolmogorov–Smirnov statistic; when $p = 1$, we are weighting the genes in S by their correlation with C normalized by the sum of the correlations over all of the genes in S. We set $p = 1$ for the examples in this paper. (See Fig. 7, which is published as supporting information on the PNAS web site.)

Estimating Significance. We assess the significance of an observed ES by comparing it with the set of scores ES_{NULL} computed with randomly assigned phenotypes.

1. Randomly assign the original phenotype labels to samples, reorder genes, and re-compute $ES(S)$.
2. Repeat step 1 for 1,000 permutations, and create a histogram of the corresponding enrichment scores ES_{NULL}.
3. Estimate nominal P value for S from ES_{NULL} by using the positive or negative portion of the distribution corresponding to the sign of the observed $ES(S)$.

Multiple Hypothesis Testing.

1. Determine $ES(S)$ for each gene set in the collection or database.
2. For each S and 1000 fixed permutations π of the phenotype labels, reorder the genes in L and determine $ES(S, \pi)$.
3. Adjust for variation in gene set size. Normalize the $ES(S, \pi)$ and the observed $ES(S)$, separately rescaling the positive and negative scores by dividing by the mean of the $ES(S, \pi)$ to yield the normalized scores $NES(S, \pi)$ and $NES(S)$ (see *Supporting Text*).
4. Compute FDR. Control the ratio of false positives to the total number of gene sets attaining a fixed level of significance separately for positive (negative) $NES(S)$ and $NES(S, \pi)$.

Create a histogram of all $NES(S, \pi)$ over all S and π. Use this null distribution to compute an FDR q value, for a given $NES(S) = NES^* \ge 0$. The *FDR* is the ratio of the percentage of all (S, π) with $NES(S, \pi) \ge 0$, whose $NES(S, \pi) \ge NES^*$, divided by the percentage of observed S with $NES(S) \ge 0$, whose $NES(S) \ge NES^*$, and similarly if $NES(S) = NES^* \le 0$.

We acknowledge discussions with or data from D. Altshuler, N. Patterson, J. Lamb, X. Xie, J.-Ph. Brunet, S. Ramaswamy, J.-P. Bourquin, B. Sellers, L. Sturla, C. Nutt, and J. C. Florez and comments from reviewers.

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270,** 467–470.
2. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., *et al.* (1996) *Nat. Biotechnol.* **14,** 1675–1680.
3. Fortunel, N. O., Otu, H. H., Ng, H. H., Chen, J., Mu, X., Chevassut, T., Li, X., Joseph, M., Bailey, C., Hatzfeld, J. A., *et al.* (2003) *Science* **302,** 393, author reply 393.
4. 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.
5. Patti, M. E., Butte, A. J., Crunkhorn, S., Cusi, K., Berria, R., Kashyap, S., Miyazaki, Y., Kohane, I., Costello, M., Saccone, R., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 8466–8471.
6. Petersen, K. F., Dufour, S., Befroy, D., Garcia, R. & Shulman, G. I. (2004) *N. Engl. J. Med.* **350,** 664–671.
7. Hollander, M. & Wolfe, D. A. (1999) *Nonparametric Statistical Methods* (Wiley, New York).
8. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. (2001) *Behav. Brain Res.* **125,** 279–284.
9. Reiner, A., Yekutieli, D. & Benjamini, Y. (2003) *Bioinformatics* **19,** 368–375.
10. Lamb, J., Ramaswamy, S., Ford, H. L., Contreras, B., Martinez, R. V., Kittrell, F. S., Zahnow, C. A., Patterson, N., Golub, T. R. & Ewen, M. E. (2003) *Cell* **114,** 323–334.
11. Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S. & Kellis, M. (2005) *Nature* **434,** 338–345.
12. Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. (2002) *Annu. Rev. Genet.* **36,** 233–278.
13. Carrel, L., Cottle, A. A., Goglin, K. C. & Willard, H. F. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 14440–14444.
14. Disteche, C. M., Filippova, G. N. & Tsuchiya, K. D. (2002) *Cytogenet. Genome Res.* **99,** 36–43.
15. Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C. & Hainaut, P. (2002) *Hum. Mutat.* **19,** 607–614.
16. Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R. & Korsmeyer, S. J. (2002) *Nat. Genet.* **30,** 41–47.
17. Zhao, N., Stoffel, A., Wang, P. W., Eisenbart, J. D., Espinosa, R., 3rd, Larson, R. A. & Le Beau, M. M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6948–6953.
18. Barbouti, A., Hoglund, M., Johansson, B., Lassen, C., Nilsson, P. G., Hagemeijer, A., Mitelman, F. & Fioretos, T. (2003) *Cancer Res.* **63,** 1202–1206.
19. Tanaka, K., Arif, M., Eguchi, M., Guo, S. X., Hayashi, Y., Asaoku, H., Kyo, T., Dohy, H. & Kamada, N. (1999) *Leukemia* **13,** 1367–1373.
20. Morelli, C., Karayianni, E., Magnanini, C., Mungall, A. J., Thorland, E., Negrini, M., Smith, D. I. & Barbanti-Brodano, G. (2002) *Oncogene* **21,** 7266–7276.
21. Mrozek, K., Heerema, N. A. & Bloomfield, C. D. (2004) *Blood Rev.* **18,** 115–136.
22. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13790–13795.
23. Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., *et al.* (2002) *Nat. Med.* **8,** 816–824.
24. Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13784–13789.
25. Smith, L. L., Coller, H. A. & Roberts, J. M. (2003) *Nat. Cell Biol.* **5,** 474–479.
26. Acker, T. & Plate, K. H. (2002) *J. Mol. Med.* **80,** 562–575.
27. Peng, T., Golub, T. R. & Sabatini, D. M. (2002) *Mol. Cell. Biol.* **22,** 5575–5584.
28. Boffa, D. J., Luan, F., Thomas, D., Yang, H., Sharma, V. K., Lagman, M. & Suthanthiran, M. (2004) *Clin. Cancer Res.* **10,** 293–300.
29. Monti, S., Savage, K. J., Kutok, J. L., Feuerhake, F., Kurtin, P., Mihm, M., Wu, B., Pasqualucci, L., Neuberg, D., Aguiar, R. C., *et al.* (2004) *Blood* **105,** 1851–1861.
30. 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.
31. 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.
32. Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C. & Conklin, B. R. (2003) *Genome Biol.* **4,** R7.
33. Zhong, S., Storch, K. F., Lipan, O., Kao, M. C., Weitz, C. J. & Wong, W. H. (2004) *Appl. Bioinformatics* **3,** 261–264.
34. Berriz, G. F., King, O. D., Bryant, B., Sander, C. & Roth, F. P. (2003) *Bioinformatics* **19,** 2502–2504.

Supporting Text

Data Sets: Description, Preprocessing and Normalization

Gene Probe to Gene Symbol Reduction. In all data sets, for each sample, the expression values of all probes for a given gene were reduced to a single value by taking the maximum expression value. By this process, the 22,283 features on the U133A chip (diabetes and gender examples) were reduced by 30% to 15,060 features, the 12,625 features on the HGU95Av2 chip (p53, leukemia, and lung Boston) were reduced by 18% to 10,104 features, and the 7,129 features on HU6800 (lung Michigan) were reduced by 10% to 6,314 features (see Table 8). Identified probe set that have no known mapping to a gene symbol were left unchanged in the data set (on average 10% of the probe sets on a chip). This probe reduction method is included in the GSEA-P JAVA package.

Description of Data Sets

Gender Data Set. This data set is unpublished (A.P.) The U133A CEL files were scaled by using the Broad Institute's RESFILEMANAGER software. Different array intensities were normalized by choosing a linear fit to the median scan (all genes). No further preprocessing was done except for gene probe reduction as described above.

P53 NCI-60 Data Set. The NCI 60 data set was downloaded from the Developmental Therapeutics Program web site (http://dtp.nci.nih.gov/mtargets/download.html). No preprocessing was done except for gene probe reduction as described above.

Leukemia Acute Lymphoid Leukemia (ALL)/Acute Myeloid Leukemia (AML) Data Set. The Leukemia data set was downloaded from ref. 1. No preprocessing was done except for gene probe reduction as described above.

Lung Cancer Data Sets. ***Michigan***. The Beer *et al.* (2) data set is available upon request.

No further preprocessing was done except for gene probe reduction as described above.

Boston. The Bhattacharjee *et al.* (3) data set was used for this study (available upon request). We extracted those lung adenocarcinomas samples for which outcome information was provided. No further preprocessing was done except for gene probe reduction as described above.

Stanford. The Stanford data set from Garber *et al.* (4) is available upon request. Missing values were replaced by zeroes. No further preprocessing was done except for gene probe reduction as described above.

Additional Detail on Gene Set Collections

Functional Sets (C2, 522 Gene Sets). The sources for sets in the C2 collection are:

1. BioCarta: www.biocarta.com.

2. Signaling pathway database: www.grt.kyushu-u.ac.jp/spad/menu.html.

3. Signaling gateway: www.signaling-gateway.org.

4. Signal transduction knowledge environment: http://stke.sciencemag.org.

5. Human protein reference database: www.hprd.org.

6. GenMAPP: www.genmapp.org.

7. Gene ontology: www.geneontology.org.

8. Sigma-Aldrich pathways: http://www.sigmaaldrich.com/Area_of_Interest/Biochemicals/Enzyme_Explorer/Key_Resources.html.

9. Gene arrays, BioScience Corp.: www.superarray.com.

10. Human cancer genome anatomy consortium: http://cgap.nci.nih.gov.

Regulatory-Motif Sets (C3, 57 Gene Sets). This catalog is based on our recent work reporting 57 commonly conserved regulatory motifs in the promoter regions of human genes (5). Some of the sites correspond to known transcription-related factors (such as SP1 and p53), whereas others are newly described. For each 8-mer motif, we identified the set of human genes that contain at least one occurrence of the motif that is conserved in the orthologous location in the human, mouse, rat, and dog genomes. These gene sets make it possible to link changes in a microarray experiment to a conserved, putative cis-regulatory element.

Neighborhood Sets (C4, 427 Gene Sets). We curated a list of 380 cancer associated genes internally and from a published cancer gene database (6). We then defined neighborhoods around these genes in four large gene expression data sets:

(*i*) Novartis normal tissue compendium (7),

(*ii*) Novartis carcinoma compendium (8),

(*iii*) Global cancer map (9), and

(*iv*) An internal large compendium of gene expression data sets, including many of our in-house Affymetrix U95 cancer samples (1,693 in all) from a variety of cancer projects representing many different tissue types, mainly primary tumors, such as prostate, breast, lung, lymphoma, leukemia, etc.

Using the profile of a given gene as a template, we ordered every other gene in the data set by its Pearson correlation coefficient. We applied a cutoff of $R \geq 0.85$ to extract correlated genes. The calculation of neighborhoods is done independently in each compendium. In this way, a given oncogene may have up to four "types" of neighborhoods according to the correlation present in each compendium. Neighborhoods with <25 genes at this threshold were omitted yielding the final 427 sets.

Additional Details on the Gene Set Enrichment Analysis (GSEA) Method. Here we elaborate on some aspects of the GSEA method that are more technical and were not described in great amount of detail in the main text due to space constraints.

Calculation of an enrichment score. Setting of the enrichment weighting exponent *p*. In the examples described in the text, and in many other examples not reported, we found that $p = 1$ (weighting by the correlation) is a very reasonable choice that allows significant gene sets with less than perfect coherence, i.e., only a subset of genes in the set are coordinately expressed, to score well. In other less common specific circumstances, one may want to use a different setting and, for this reason, the GSEA-P program accepts *p* as an input parameter. For example, if one is interested in penalizing sets for lack of coherence or to discover sets with any type of nonrandom distribution of tags, a value $p < 1$ might be appropriate. On the other hand, if one uses sets with large number of genes and only a small subset of those is expected to be coherent, then one could consider using $p > 1$. Our recommendation is to use $p = 1$ and use other settings only if you are very experienced with the method and its behavior.

Benefits of Weighting by Gene Correlation. Most gene sets show some amount of coherent behavior but are far from being perfectly coherent. For example in Fig. 7, we show the enrichment plot for the set of genes up-regulated by p53 in the p53 wild-type phenotype. This set is one of those that is significantly enriched by using the current GSEA method. However, if we use the original constant weight for GSEA analysis, this set is not significant. This failure to affirm significance is an issue is a problem because

we would expect such a set to be enriched for the p53 wild-type phenotype. From the figure, we can see that the 40 genes in the set are not uniformly coherent but rather split into two coexpressed groups with some additional scatter. The use of equal weighting tends to overpenalize this lack of coherence and does not produce a significant enrichment score (*ES*) for this gene set, even though a significant subset of its genes are near the top of the list.

Multiple Hypothesis Testing

Adjusting for Variation in Gene Set Size. As described in *Appendix*, when adjusting for variation in gene set size, we normalize the *ES(S,* π) for a given *S*, separately rescaling the positive and negative scores by dividing by their mean value, yielding *NES(S,* π) and *NES(S)* (normalized scores, *NES*).

This gene set size normalization procedure appropriately aligns the null distributions for different gene sets and is motivated by the asymptotic multiplicative scaling of the Kolmogorov-Smirnov distribution as a function of size (10). Here, we will make a brief digression to elaborate on this subject.

The analytic form of the Kolmogorov-Smirnov distribution scaling with gene set size can be derived from the expectation value of the approximated distribution function of the enrichment statistic:

$$\Pr\left(ES\left(N,N_H\right)<\lambda\right)=\sum_{k=-\infty}^{\infty}\left(-1\right)^k\exp\left(-2k^2\lambda^2 n\right),\ n=\frac{(N-N_H)N_H}{N}, \quad \textbf{[1]}$$

where λ is the enrichment score, N is the number of genes in the gene list, and N_H the number of genes in the gene set. The number of terms required for the above series to converge depends on λ. As λ approaches zero, more terms are required. From the above equation, we can compute the following density function for the enrichment statistic

$$\rho(\lambda) = 4\sum_{k=-\infty}^{\infty}(-1)^{k+1}k^2 n\lambda \exp\left(-2k^2\lambda^2 n\right). \quad \textbf{[2]}$$

Notice the multiplicative scaling of the distribution with n, and for large gene lists ($N >> N_H$) with N_H.

The average enrichment score is simply the expectation (integral from $\lambda = 0$ to 1), with respect to the above density:

$$\overline{ES} = \mathrm{E}_{\rho(\lambda)}\, ES(N, N_H) = \int_{\lambda=0}^{1} \lambda\rho(\lambda)\, d\lambda$$
$$= 4\sum_{k=-\infty, k\neq 0}^{\infty}(-1)^{k+1}\left(\tfrac{1}{4}\exp\left(-2k^2 n\right) - \frac{\sqrt{2\pi}}{16}\frac{\mathrm{erf}\left(\sqrt{2n}k\right)}{k\sqrt{n}}\right). \quad \textbf{[3]}$$

where erf is the “error function” (integral of the normal distribution).

The mean values of the null distribution of enrichment scores computed with this approximation are quite consistent with our actual empirical results when using GSEA unweighted enrichment scores ($p = 0$). Therefore if we were only performing unweighted GSEA and permuting the genes, we could analytically compute the normalization factor by using the equation above. However, our standard practice is to use weighting and to permute the phenotype labels; therefore, this expression is not entirely accurate.

For example when using GSEA weighted scores ($p = 1$), the empirical mean values are ≈5 times smaller. This expected reduction in “effective” gene set size is the direct effect of gene-gene correlations. Notice that these correlations are preserved by the phenotype label permutation and are also relevant when using the correlation profiles as part of the weighted GSEA enrichment score calculation. Despite the change in the mean, the shape of the distribution is still very much the same, and multiplicative scaling works well empirically for the gene set size normalization.

Computing Significance By Using Positive or Negative Sides of the Observed and Null Bimodal ES Distributions: As mentioned in the main text, the use of a weighted enrichments score helps make the current GSEA method more sensitive and eliminates some of the limitations of the original GSEA method; however, it also makes more apparent any lack of symmetry in the distribution of observed *ES* values. This intrinsic asymmetry can be due to class specific biases either in the gene correlations or in the population of the gene set collection itself (Fig. 4). Specifically, many more genes may be highly correlated with one phenotype, or the collection of gene sets may contain more that are related to one of the two phenotypes. On the other hand, constructing the null by using random phenotype assignments tends to produce a more symmetric distribution that may not exactly coincide with the bulk, nonextreme part of the distribution of the observed values. To address this issue, we determine significance and adjust for multiple hypotheses testing by independently using the positive and negative sides of the observed and null bimodal ES distributions. In this way, the significance tests [nominal *P* value, familywise-error rate (FWER), and false discovery rate (FDR)] are single tail tests on the appropriate (positive/negative) side of the null distribution.

FWER. The use of FWER, which controls the probability of a false positive, to correct for multiple hypothesis testing (MHT) in the original GSEA method is overly conservative and often yields no statistically significant gene sets. For example, the analysis results by using the original GSEA method do not produce any significant set (FWER < 0.05) on either side in the Gender, Leukemia, and p53 examples. Nonetheless, the GSEA-P program also computes the familywise error by creating a histogram of the maximum $NES(S, \pi)$ over all S for each π by using the positive or negative values corresponding to the sign of the observed $NES(S)$. This null distribution is then used to compute an FWER p value.

Description of GSEA Output. The output of the GSEA-P software includes a list of the gene sets sorted by their *NES* values along with their nominal and FWER p values and their FDR q values.

The GSEA-P R and JAVA programs compute several additional statistics that may be useful to the advanced user:

***Tag %*:** The percentage of gene tags before (for positive *ES*) or after (for negative *ES*) the peak in the running enrichment score *S*. The larger the percentage, the more tags in the gene set contribute to the final enrichment score.

***Gene %*:** The percentage of genes in the gene list *L* before (for positive *ES*) or after (for negative *ES*) the peak in the running enrichment score, thus it gives an indication of where in the list the enrichment score is attained.

***Signal strength*:** The enrichment signal strength that combines the two previous statistics: $(Tag\ \%) \times (1 - Gene\ \%) \times (N / (N - N_h)$, where *n* equals the number of genes in the list and N_h is the number of genes in the gene set. The larger this quantity, the more enriched the gene set is as a whole. If the gene set is entirely within the first N_h positions in the list, then the signal strength is maximal or 1. If the gene set is spread throughout the list, then the signal strength decreases toward 0.

***FDR (median)*:** An additional FDR *q* value was computed by using a median null distribution. These values are, in general, more optimistic than the standard FDR *q* values as the median null is a representative of the typical random permutation null rather than the extremes. For this reason, we do not recommend it for common use. However, the FDR median is sometimes useful as a binary indicator function (zero vs. nonzero). When it is zero, it indicates that for those extreme *NES* values the observed scores are larger than the values obtained by at least half of the random permutations. One advantage of selecting gene sets in this manner (FDR median = 0) is that a predefined threshold is not required. In practice the gene sets selected in this way appear to be roughly the same as those for which the regular FDR is <0.25. For example, in the Leukemia ALL/AML example, the FDR median is 0 for the five top scoring sets (four of which have FDR < 0.25).

***glob.p.val*:** A global nominal *P* value for each gene set's NES estimated by the percentage of all (S, π) with $NES(S, \pi) \geq NES(S)$. Theoretically, for a given level of significance (e.g., 0.05), this quantity measures whether the shift of the tail of the distribution of observed values is extreme enough to declare the observed distribution as different from the null. In principle, it allows us to compute a quantitative measure of whether there is any enrichment in the data set with respect to the given database of gene sets. In practice, this quantity behaves in a somewhat noisy way because of the sparseness in the tail of the observed distribution.

One Set of Global Reports and Plots. They include the scores and significance estimates for each gene set, the gene list correlation profile, the global observed and null densities, and a heat map for the sorted data set.

A Variable Number of Specific Gene Set Reports and Plots (One for Each Gene Set). These reports include a list of the members of the set and the leading-edge, a gene set running enrichment "mountain" plot, the gene set null distribution and a heat map for genes in the gene set.

The format (columns) for the global result files is as follows: GS, gene set name; size, number of genes in the set; source, set definition or source; *ES*, enrichment score; *NES*, normalized (multiplicative rescaling) enrichment score; NOM *p* val, Nominal *p* value (from the null distribution of the gene set); FDR *q* val, false discovery rate q values; FWER *p* val, familywise error rate *p* values; tag %, percent of gene set before running enrichment peak; gene %, percent of gene list before running enrichment peak; signal, enrichment signal strength; FDR (median), FDR *q* values from the median of the null distributions; glob.p.val, *p* value by using a global statistic (number of sets above the given set's *NES*).

The rows are sorted by the *NES* values (from maximum positive or negative *NES* to minimum).

The format (columns) for the individual gene set result files contains the following information for each gene in the set: Probe_ID, the gene name or accession number in the data set; symbol, gene symbol from the gene annotation file; Desc, gene description (title) from the gene annotation file; list loc, location of the gene in the sorted gene list; S2N, signal-to-noise ratio (correlation) of the gene in the gene list; RES, value of the running enrichment score at the gene location; core_enrichment, yes or no variable specifying if the gene is in the leading-edge subset.

The rows are sorted by the gene location in the gene list.

Post-GSEA Analysis: Leading-Edge Subset Similarity, Clustering, and Assignment. In analyzing the top scoring gene sets resulting from GSEA, we may wish to determine whether their GSEA signal derives from a common subset of genes. These shared subsets tells us whether we should interpret the sets as representatives of independent processes, or if, in fact, they result from the same common mechanism. If we find that this subset of genes behaves similarly and coherently, we may wish to treat it as a new gene set in one of our collections.

To make the discovery of such common, overlapping signals with the leading-edge subsets of high-scoring gene sets, we have created software that reads the GSEA results and creates several postanalysis reports and visualizations. The software performs the following three basic types of analyses:

(*i*) Creates a similarity matrix heat map that shows at a glance whether leading-edge subsets of two gene sets are highly overlapping.

(*ii*) Creates an assignment matrix of gene sets vs. leading-edge genes for each phenotype. This binary matrix shows explicitly the membership of each gene in each high-scoring gene set and the overlaps between the gene sets.

(*iii*) Performs a hierarchical clustering (by using average linkage) and re-sorts the genes and gene sets in the assignment matrix according to their similarity to create clustered assignment matrices for each phenotype. This clustering helps to uncover common occurrences of the same leading-edge genes in several gene sets.

As described in the paper, we used this program to study the top scoring gene sets enriched in the p53 mutant cancer cell lines (see Fig. 3).

This type of analysis helps in the interpretation of GSEA results and the identification of leading-edge overlaps between gene sets that are responsible for high enrichment scores. If applied systematically, it can also provide a method for refining genes sets and creating new ones.

Original GSEA Method from Mootha *et al.* Here we described the original GSEA method as defined in Mootha *et al.* (11).

Calculate enrichment. We set the constant step size of the walk, so that it begins and ends with 0, and the area under the running sum is fixed to account for variations in gene set size. We walk down the list *L*, incrementing the running sum statistic by $\sqrt{(N - N_h)/N_h}$ when we encounter a gene in *S* and decrementing by $\sqrt{N_h/(N - N_h)}$ if the gene is not in *S*, where N is the number of genes in the list *L*, and N_h is the number of genes in the gene set *S*. The maximum deviation from zero is the *ES* for the gene set *S*, and corresponds to a standard Kolmogorov-Smirnov statistic (12).

Determine the significance of ES. We permuted the phenotype labels and recomputed the *ES* of a gene set to generate a null distribution of *ES*. Using this null, we computed an empirical, nominal *p* value for the observed *ES*.

Adjust for MHT. When scoring multiple gene sets we constructed a null distribution to estimate the FWER by constructing a histogram of the maximum *ES* score achieved by

any gene set for a given permutation of the phenotype labels. The FWER provides a very conservative correction, which controls the probability of even a single false positive.

Notice that except for the normalization procedure (and the use of FDR instead of FWER), the current GSEA method with $p = 0$ is quite similar to this original GSEA method.

GSEA-P R Program. The R scripts and data that produced the results and figures in this paper are available upon request.

1. Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R. & Korsmeyer, S. J. (2002) *Nat. Genet.* **30,** 41–47.

2. Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., *et al.* (2002) *Nat. Med.* **8,** 816–824.

3. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13790–13795.

4. Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13784–13789.

5. Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S. & Kellis, M. (2005) *Nature* **434,** 338–345.

6. Brentani, H., Caballero, O. L., Camargo, A. A., da Silva, A. M., da Silva, W. A., Jr., Dias Neto, E., Grivet, M., Gruber, A., Guimaraes, P. E., Hide, W., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 13418–13423.

7. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101,** 6062–6067.

8. Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F., Jr., *et al.* (2001) *Cancer Res.* **61,** 7388–7393.

9. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 15149–15154.

10. von Mises, R. (1964) *Mathematical Theory of Probability and Statistics* (Academic, New York).

11. 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.

12. Hollander, M. & Wolfe, D. A. (1999) *Nonparametric Statistical Methods* (Wiley, New York).

Table 3. Molecular Signature Database (MSigDB) collections

Database	Total number of sets	Number of sets smaller than 15	Number of sets greater than 500	Number of sets used	Percent used
C1	319	74	10	235	74
C2	522	194	0	328	63
C3	57	0	0	57	100
C4	427	0	0	427	100

Listed are the number of gene sets in each collection, the numbers of sets that pass the size thresholds (min = 15, max = 500), and the final number of sets used in the examples in the text.

Table 5. Functional autosomal gene set enrichment with respect to gender

Gene set	Source	*ES*	*NES*	*Nominal p-value*	*FDR*
Data set: Lymphoblast cell lines					
Enriched in males					
C2:Testis expressed autosomal genes	Experimental GNF	0.559	1.724	0.001	0.181
Enriched in females					
C2:Female reproductive tissue expressed autosomal genes	Experimental GNF	-0.457	-1.830	0.004	0.163

This table shows the GSEA results for the gender data set by using the functional collection C2 after restricting the gene set membership to autosomal genes. *ES*, enrichment score; *NES*, normalized enrichment score.

Table 6. Single gene overlaps in lung cancer studies

Michigan / Boston (12)	Michigan / Stanford (8)	Stanford / Boston (4)
KRT7	**KRT7**	**KRT7**
BZW1	CSNK1E	CASP4
CASP4	GALNT3	GOSR1
CSNK1E	HIP2	PAICS
ENO2	ITGA2	
FADD	NP	
KRT18	NPAS2	
KRT19	PAICS	
LAMB3		
P4HA1		
PFKP		
TUBA1		

Pairwise overlaps are shown between the top 100 genes correlated with poor outcome in the Michigan, Boston, and Stanford data sets as depicted in Fig. 5. Pairwise overlap is determined by using genes that appear on the technology platforms of both studies. Restricting to genes on all 3 platforms would reduce the gene space by 50% in the Michigan study and by 70% in the Boston and Stanford studies.

Table 8. Probe set to gene ID reduction

Number of probe set ids per gene	Number of genes	Percent of genes
1	10,553	70.07
2	2,758	18.31
3	1,136	7.54
4	407	2.70
5	128	0.85
6	46	0.31
7	13	0.09
8	9	0.06
9	5	0.03
10	3	0.02
11	1	0.01
12	0	0.00
13	1	0.01
Total	15,060	

HGU133A

Number of probe set ids per gene	Number of genes	Percent of genes
1	8,276	81.91
2	1,326	13.12
3	381	3.77
4	83	0.82
5	18	0.18
6	11	0.11
7	6	0.06
8	3	0.03
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
Total	10,104	

HGU95AV2

Number of probe set ids per gene	Number of genes	Percent of genes
1	5,670	89.8
2	516	8.17
3	102	1.62
4	18	0.29
5	2	0.03
6	4	0.06
7	1	0.02
8	1	0.02
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
Total	6,314	

HU6800

The distribution of probe sets per gene in the three Affymetrix chip types used on the data sets in the paper are shown. The data displayed is binned by the number of probes per gene. The majority of the overrepresentation arises from two or three probes per gene. In our analyses, we chose the maximally expressed probe as the single representative of the corresponding gene.