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 This data set is unpublished (A.P.) The U133A CEL files were scaled by using the Broad Institute's RESFILEMANAGER software.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).