

Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm/k-nearest Neighbor Method

Leping Li^{1,*}, Thomas A. Darden², Clarice R. Weinberg³, A. J. Levine⁴ and Lee G. Pedersen^{2,5}

¹ Exposure Assessment Branch, Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, West Virginia 26505, USA

² Laboratory of Structural Biology and ³ Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA

⁴ President's Office, Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

⁵ Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290, USA

*Correspondence should be addressed to L.L. (voice: 919-541-5168, fax: 919-541-4311, email:

Li3@niehs.nih.gov

Keywords: pattern recognition, gene selection, high-dimensional, microarray

A running title: sample classification using GA/KNN

Abstract

Recent tools that analyze microarray expression data have exploited correlation-based approaches such as clustering analysis. We describe a new method for assessing the importance of genes for sample classification based on expression data. Our approach combines a genetic algorithm (GA) and the k-nearest neighbor (KNN) method to identify genes that jointly can discriminate between two types of samples (e.g. normal vs. tumor). First, many such subsets of differentially expressed genes are obtained independently using the GA. Then, the overall frequency with which genes were selected is used to deduce the relative importance of genes for sample classification. Sample heterogeneity is accommodated; that is, the method should be robust against the existence of distinct subtypes. We applied GA/KNN to expression data from normal versus tumor tissue from human colon. Two distinct clusters were observed when the 50 most frequently selected genes were used to classify all of the samples in the data sets studied and the majority of samples were classified correctly. Identification of a set of differentially expressed genes could aid in tumor diagnosis and could also serve to identify disease subtypes that may benefit from distinct clinical approaches to treatment.

Introduction

Recent advances in microarray technology have made it possible to study the expression patterns of thousands of genes in parallel (for reviews see refs 1-4). Microarrays have become valuable tools for studying the gene expression patterns of normal and diseased tissues [5-7] the genome-wide patterns of gene expression of microorganisms under different conditions [8,9], changes in gene expression patterns of cells in response to environmental and genotypic changes [10], and changes in gene expression patterns of cells as a function of time [11]. For instance, Alon *et al.* [6] used oligonucleotide arrays to study the expression patterns of tumor and normal colon tissue samples. In their studies, the gene expression patterns of 40 colon tumor tissue samples and 22 normal colon tissue samples were analyzed with an Affymetrix oligonucleotide array [4] complementary to more than 6,500 human genes and expressed sequence tags (ESTs). Gene correlation with tissue classification as well as discrimination between normal and tumor tissue samples was obtained using a cluster analysis.

While high-throughput technology has significantly accelerated the rate at which biological information is acquired, tools that can successfully mine the resulting large data sets are needed. Currently, the methods commonly applied to microarray data analysis have been correlation-based approaches such as

cluster analysis [6,12]. Cluster analysis groups genes that are similar in patterns of expression. Clustered genes are likely to be functionally linked. Likewise, this approach provides valuable information about gene interactions and gene relationships, from which the functions of specific genes and their cellular locations and roles in specific pathways may be suggested [13,14]. Although correlation-based approaches have been widely applied in analyzing the patterns of gene expression, they may not fully extract the information from data corrupted by high-dimensional noise. This is important because samples may be incorrectly classified if the noise from the genes that are irrelevant is not sufficiently reduced.

Methods for selecting informative genes for sample classification have been recently proposed [15,16]. Golub *et al.* [15] developed a neighborhood analysis approach to obtain a subset of genes that discriminate between the acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Genes whose expression levels differ significantly in ALL and AML were identified. These genes were subsequently used to predict the class membership of new leukemia cases. Recently, Ben-Dor *et al.* [16] applied a boosting technique [17] to search for a threshold (expression level) for each gene that would maximally distinguish between two types of samples. Those that gave the smallest classification errors were taken as the relevant genes. Essentially, both approaches [15,16] are univariate approaches, that is, samples were compared in each single gene dimension. It is not clear that reduction of the gene dimensions from thousands to one retains the essential information necessary for sample distinction. In addition, both methods implicitly assume that the relevant genes are similarly expressed among samples of each type (e.g. tumors). This could be problematic when subtypes exist so that the relevant genes are not uniformly expressed.

Herein we present a multivariate approach that compares samples in a multi-gene dimension using a nonparametric pattern recognition approach, the k-nearest neighbor method (KNN) [18]. A sample is classified based on the class membership of its nearest neighbors in the gene space. The dimensionality of the gene subspace is arbitrarily set to 50. Intuitively, samples (objects) become more distinct (dissimilar) when more genes are compared. On the other hand, too many dimensions may contribute noise to the system. In addition, the calculation becomes computationally expensive as the number of dimensions (genes) increases. A 50-dimension gene subspace is a reasonable compromise.

We began by using the “genetic algorithm” to select many subsets of 50 genes that can potentially discriminate between tumor and normal tissue samples in the Alon *et al.* data set [6]. When a large number of such subsets of genes were obtained, the frequency with which genes were selected was assessed through statistical analysis. The selection frequency should correlate with the relative predictive importance of genes for sample classification: the most frequently selected genes should be jointly discriminative, and therefore should include genes that are differentially expressed. The most frequently selected genes could subsequently be used to classify new samples; that is, potentially be used for tumor diagnosis.

Comparing all subsets of genes is not a feasible approach. For instance, the number of ways to select 50 genes from 2000 is approximately 10^{100} . It is not possible to examine all the combinations directly. An efficient method is needed to sample from fewer combinations to find the optimal or near optimal solutions. Although many optimization methods may be in principle appropriate for this task, genetic algorithms (GAs) provide a general purpose, stochastic search methodology. GA has been used in a variety of combinatorial problems involving high dimensional spaces [19,20]. In this study, we applied GA in combination with the KNN method to identify the many subsets of genes that can discriminate tumor from normal tissue samples. We tested our method on the data set of Alon *et al.* [6]. In the data set, there are 62 colon tissue samples (40 tumor and 22 normal), of which each contains the expression levels of 2,000 genes/ESTs.

Methods

Data set. The original gene expression data were downloaded from the web. The data contain the expression levels of 2000 genes across the 62 samples, of which 40 are tumor tissue and 22 normal tissue [6]. The data had previously been filtered from oligonucleotide microarray studies of the expression levels of 3200 full-length human cDNA and 3,400 ESTs that have some similarity to other eukaryotic genes using

Affymetrix Oligonucleotide Arrays [4]. The data set was divided into a training set (the first 42 samples) and a test set (20 samples). The numbers of tumor and normal tissue samples are 28 and 14 in the training set and 12 and 8 in the test set, respectively. This 2-to-1 split reflects the more complex role played by the training set in selecting the genes from the 2000.

K-nearest neighbors (KNN). In the KNN method [18], one computes the distance between a sample, represented by its pattern vector V_m , and each of the pattern vectors of the training set:

$V_m = (g_{1m} \dots g_{im} \dots g_{nm})$, where n is the number of genes in the vector (set to 50 arbitrarily); g_{im} is the expression level (\log_{10} transformed) of the i th gene in the m th sample; $m = 1, \dots, M$.

Each sample is classified according to the class membership of its k nearest neighbors (provided they agree), as determined by the Euclidean distance in 50-dimensional space. If the k nearest neighbors are not of the same class, the sample is considered unclassifiable. Small values of 3 or 5 have been alleged to provide good classification [18]. We arbitrarily set k to be 3, large enough to form tight clusters even if there are subtypes and the sample size is limited. A larger k would allow less flexibility in detecting subclusters and also increase required computing time. Unlike the classic KNN for which class membership is determined by majority vote of the k -nearest neighbors [18], we require a unanimous decision. If all of the k -nearest neighbors of a sample are tumors, the sample is classified as tumor and conversely. A sample remains unclassified if its three nearest neighbors are not of the same class.

Genetic algorithm (GA). GA, first described by John Holland [21], mimics natural evolution and selection. In biological systems, genetic information that determines the individuality of an organism is stored in chromosomes. Chromosomes are replicated and passed onto the next generation with selection depending on fitness. Genetic information can, however, also be altered through genetic operations such as mutation and crossover. In GAs, each “chromosome” is a set of genes, which constitutes a candidate solution to the discrimination problem. A population of “chromosomes” is used. The passage of each “chromosome” to the next generation is determined by its relative fitness, i.e. the closeness of its properties to those desired. Random combinations and/or changes of the transmitted “chromosomes” produce variations in the next generation of “offspring”. The better the fitness (correspondence with desired properties), the greater the chance of being selected for transmission. Through evolution through many generations, optimal or near optimal solutions are obtained. There are four major components of GA: chromosome, fitness, selection, and mutation.

In this study, a GA was used to identify 50 genes that can correctly classify all of the training set samples. For some applications a less strict criterion may be preferred, e.g. allowing one or two misclassifications. A schematic diagram of the GA/KNN procedure is shown in Fig. (1).

chromosomes

Each “chromosome” consisted of 50 selected genes. Initially, a diverse set of chromosomes (called a “population”) was generated. Each chromosome was generated by randomly selecting 50 distinct genes from the 2,000 gene pool. Multiple populations were generated, called sub-populations or “niches” in GA. For a typical run, 10 niches were separately evolved where each contained 150 chromosomes. Each niche evolves independently, except that at each generation, the best chromosomes identified, one from each niche, were combined and used to replace the 10 least fit chromosomes in *each* niche in the next generation (Fig. 1). This strategy preserves the best chromosomes at each generation and also shortens the search time [22].

fitness function

For each selected set of 50 genes (a chromosome), the class memberships of the three nearest (training set) neighbors were compared to that of each particular sample. If all 4 class memberships agreed,

a score of 1 was assigned to the particular sample. These scores were summed across samples to assess the goodness of classification of each particular set of 50 genes, forming a sum we refer to as R^2 . The maximum R^2 thus corresponds to the total number of samples in the training set, and R^2/M gives the proportion correctly classified, where M is the number of samples in training. Each chromosome in the population was assigned a fitness score based on its ability to classify the samples in the training set based on R^2 , as defined above.

selection

If the initial population of chromosomes does not include any that achieve the maximal R^2 , a new set of chromosomes is generated to form a second generation. Selection of chromosomes for the next generation is based on the *survival-of-the-fittest* principle. The single best chromosome from each niche is entered into the respective subsequent niche deterministically and the remaining 149 positions are filled based on the fitness values (probabilistically).

mutation

Mutation is employed to enhance evolvability by introducing new genes into the chromosomes. Each chromosome was selected from the parent niche according to a probability proportional to its fitness rank. Once a chromosome is chosen for transmission, between 1 and 5 of its genes are randomly selected for mutation. The number of mutations is assigned randomly, with probabilities, 0.53125, 0.25, 0.125, 0.0625, and 0.03125 respectively. In this way, a single replacement is given the highest probability while simultaneous multiple replacement has low probability. This strategy prevents the search from behaving as a random walk as it would if many new genes were introduced at each generation. Once the number of genes in the chromosome to be replaced has been determined, these genes are randomly selected and replaced randomly from the genes not already in the chromosome.

The above procedure was repeated until a maximum in the cross-validation R^2 (i.e. M) was found in any of the 10 niche runs (typically, in 10-50 generations). The selected chromosome was saved. The whole population was independently regenerated for each niche and the process was repeated. The process was terminated when an arbitrary large number of chromosomes of genes were obtained. The number of selected chromosomes was 6,348 for the colon data.

Principal component (PC) analysis. For visualization, a principal component (PC) analysis [23] was applied to the variance-covariance matrix of the data (62 samples, \log_{10} transformed) using the method of single value decomposition (SVD) from Numerical Recipes (Cambridge, MA) to extract the eigenvalues. The principal components were obtained by projecting the original data points (\log_{10} transformed) onto the eigenvectors.

Results

A total of 6,348 subsets of 50 genes that potentially discriminate between the normal and tumor samples were obtained based on the training set samples. The frequency with which genes were selected was then analyzed. The 50 most frequently selected genes were subsequently used to classify samples in the test set.

Gene Selection

The statistical z -score with which each of the 2,000 genes was chosen from the 6,348 solutions is shown in Fig. (2). To determine if GA has adequately sampled the solution space, the 6,348 solutions were divided into two equal-size groups and their frequency distributions were compared. Nearly identical patterns were observed. Invariably, several genes were being selected more frequently than others while a few were never selected. The statistical significance analysis (Fig. 2) indicates that selection is not random.

Although many genes were selected with significantly high z -scores, only the 50 most frequently selected genes are listed in Table 1. The complete list of the 2,000 genes based on frequency rank is available (<http://dir.niehs.nih.gov/microarray/datamining/>).

The human monocyte-derived neutrophil-activating protein (MONAP) gene was most frequently selected. This gene is significantly up-regulated in the tumor tissue samples compared to the normal tissue samples (Student *t*-test [24], $P < 0.0001$). Recent studies have demonstrated that the expression level of MONAP, also called interleukin-8 (IL-8), directly correlates with the progression of several human cancers [25,26]. It is believed that over-expression of IL-8 may play an important role in tumor angiogenesis and aggression [25,26].

Among the top 50 genes, 5 are antigen or antigen-related genes. These genes are the T-cell acute lymphoblastic leukemia associated antigen 1, HLA class II histocompatibility antigen γ chain precursor (*H. sapiens*), LCA-homolog - LAR protein (leukocyte antigen related), transmembrane carcinoembryonic antigen BGP α (formerly TM1-CEA) and transmembrane carcinoembryonic antigen BGPC (formerly TM3-CEA). Another antigen gene, which nearly made the top 50 list, is the leukocyte antigen CD37 (*H. sapiens*) (rank no. 51).

Several putative tumor suppressor genes were also among the top 50 genes (Table 1). They are human *MXI1* (MAX-interacting protein 1), gelsolin precursor, p-cadherin, and tropomyosin. All except p-cadherin were significantly down-regulated in tumors compared to normal samples. *MXI1* is believed to be a member of the MYC family of transcription factors that negatively regulates MYC function [27]. Correlation between *MXI1* genetic instability and tumors has been reported [28,29]. Gelsolin is an actin filament regulatory protein that plays an important role in maintaining the integrity of cell cytoskeleton [30]. It has been suggested as a tumor suppressor because its expression is evidently reduced or lost in several tumors including breast [31], ovarian adenocarcinomas [32], and prostate [33]. P-cadherin belongs to a family of cell-cell adhesion molecules that are essential to embryonic development, maintenance of tissue integrity and tumorigenesis [34]. Among the four putative suppressor genes we selected, it was the only one that was highly expressed in tumors compared to normal samples. Over-expression of p-cadherin in breast carcinoma has been strongly associated with poor survival prognosis [35]. Tropomyosin is another suppressor protein that suppresses cell malignant transformation [36]. In addition, the vasoactive intestinal peptide (VIP) gene was also frequently selected. It has been shown that VIP inhibits the proliferation of human colonic cancer cells line HT29 [37] and other cancer cell lines [38]. Thus, it is not surprising that the *VIP* gene was substantially down-regulated in the tumor tissue samples compared to the normal tissue samples. Several cell adhesion and skeletal related genes were also frequently selected.

Among the 2,000 genes, 12 were controls, labeled as HSAC07, UMGAP and I. One would expect these genes not to be frequently selected. In fact, four (labeled as I) are among the 9 genes that were never selected in 6,348 solutions. The other 8 were selected with frequencies near random (from 0.015 to 0.046). Multiple copies of genes on the chip provide another kind of control, and these multiples were selected with similar frequencies (data not shown).

Validation

To test the predictive strength of our selected 50 genes on a test set of specimens, each of the 20 test set samples was classified according to the class memberships of its three nearest training set neighbors using the most frequently selected genes (Table 2). When only the top most gene was used to classify the test set samples, 7 were incorrectly classified while 2 remained unclassifiable. It appears that the prediction stabilizes when as few as 25 and up to 110 top genes are used. When up to 120 top genes were used, a similar result was obtained except that T30 became unclassifiable. As more genes were included, contaminating the system with high-dimensional noise, the number of unclassifiable samples increased. In fact, when all 2,000 genes were used, 8 out of 20 samples became unclassifiable. This result emphasizes that not all expression data are relevant to the discrimination between the normal and tumor samples. For those that were classified, two normals (N34 and N36) and three tumors (T30, T33, and T36) were incorrect. In all cases, the two normal samples N34 and N36 were again classified as tumors. As expected, the least frequently selected 100 genes were unable to discriminate between tumor and normal samples (Table 2).

Data Display

Principal component (PC) analysis [23] is a statistical method, often called eigenvector analysis, eigenvector decomposition or Karhunen-Loeve expansion. In PC, most of the variation in data is summarized by projection onto a few orthogonal principal components. More specifically, the first PC is the major axis for the “shape” of the data points. Hence, the first few PCs explain most of the variance in the data. A plot of the first two PCs often reveals patterns in the data. Here it was used to project samples of high dimensions onto a two-dimension plot for visual display. Other display or post processing techniques such as the cluster analysis [6,12] could be used (Fig. 3).

We applied PC analysis to all of the 62 samples in the 2000-dimensional gene space. Several PC were obtained. The first two PCs represent approximately 60% of the variance of the data set. Noticeably, no separation between tumor and normal samples is apparent (Fig. 4a). In contrast, two distinct clusters emerge when only the 50 most frequently selected genes are used (Fig. 4b). A similar pattern was observed when the 100 or 200 most frequently selected genes were used (data not shown). Again, two normal samples (N34 and N36) were in the cluster of tumors while three tumor samples (T30, T33 and T36) were among the normal samples. Clustering analysis using all 2000 genes showed that N34 is positional among a cluster of tumors while N36 is adjacent to one tumor (T2) in the cluster diagram [6]. All the three tumor samples (T30, T33 and T36) are in the cluster of normal samples in the cluster diagram [6]. Thus, results based on GA/KNN (with or without subsequent PC analysis) are consistent with some of those provided by cluster analysis. Together, the results suggest that samples T30, T33, N34, N36, and T36, but especially the latter four are anomalies. Communication with Dr. Uri Alon, the first author of reference 6, sheds some light on the question. Here we quote “It appears that tissue samples are heterogeneous, that is, they have a mixture of epithelial cells (which are the cancerous cells in the case of tumor samples) and other tissues such as muscle cells which are not cancerous. Therefore, it is plausible that much of the tumor/normal classification found by clustering is actually due to differences in the fraction of epithelial cells in the sample. When one makes a crude estimate of the amount of, say, muscle cells in each sample by taking the average intensity of ‘muscle genes’, one sees that the tumor samples contain less muscle than the normal samples. The exceptions to this are precisely the outliers; they seem to have a different ratio of tissue compositions. The 3 tumors have a high muscle content, and the 2 normals have a low muscle content”. This suggests that sample contamination can distort the classification process.

As expected, no distinctive clusters were observed when PC analysis was applied to 50 randomly selected genes (Fig. 4c) or the 100 least frequently selected genes (Fig. 4d).

Discussion

Method development for analyzing gene expression data is still in its infancy. Nonetheless, different approaches are emerging [6,12,15,16,39-41]. Recently, Ben-Dor *et al.* [16] applied several methods to sample classification including support vector machines (SVM) [42]. In SVMs, one seeks a hyperplane that can separate two groups of points (e.g. normal vs. tumor samples) and maximizes the minimum distance of the closest points to the hyperplane. SVM has also been applied to gene classification [39]. While most methods utilize all the expression data (after filtration) in the analysis, methods that use a subset of relevant genes have been reported [15,16]. For instance, Golub *et al.* [15] applied neighborhood analysis to identify a subset of genes that discriminate between the two types of leukemia AML and ALL, using a separation measure similar to the *t*-statistic. The 50 genes that best distinguish AML from ALL using 38 training set samples were taken as the informative genes. Subsequent classification using 50 informative genes correctly predicted 29 of 34 test set samples with high confidence. Ben-Dor *et al.* [16] also applied a boosting technique [17] to search for a threshold (expression level) for each gene that would maximally discriminate between two types of samples. Those that gave the smallest classification errors were taken as the relevant genes. The method was applied to the same colon data set (we have used) and an ovarian set. All samples were used to obtain the relevant genes using the *leave-one-out cross-validation* procedure as a measure of prediction strength (no test set was used). Although differing in technical details, both approaches (Golub *et al.* [15] and Ben-Dor *et al.* [16]) examine one gene at a time (univariate). Furthermore, both approaches

[15,16] implicitly assume that genes are similarly expressed within each type of samples. This could be problematic when subtypes exist so that the relevant genes are not uniformly expressed in the group. On the other hand, the GA/KNN approach is a multivariate approach (samples are compared in multi-gene dimensions). A sample is classified based on the class memberships of its nearest neighbors. Therefore, subclusters among samples of a given class are accommodated. We found that with the colon data no single gene was capable of discriminating between all the normal and tumor samples (using KNN, $k = 3$ and consensus rule) while many subsets of combinations of genes can do this. This observation was also borne out for other array data that we analyzed including leukemia (<http://www.genome.wi.mit.edu/MPR>) and lymphoma data (<http://lmpp.nih.gov/lymphoma>).

We have found that combinations of genes, rather than individual dominant genes, can be most discriminating for sample classification. It appears that the discriminative effect of certain genes is synergistic. For instance, the frequency of co-occurrence of three genes (P_{ABC}) p145TRK-B (A), Tropomyosin (B), and Metallothionein-II (C), is approximately 6 times higher than would be predicted based on their individual frequencies ($P_A \times P_B \times P_C$). Many such examples exist. The GA/KNN method, by selecting sets of genes based on their joint ability to discriminate, can in theory identify genes that are important jointly, but which would not appear to discriminate individually. For example, 3 genes that we highly selected (Glucocorticoid receptor, p145TRK, and 60S acid ribosomal protein P2) each had a t -statistic that was unimpressive (1.38, 0.44, and 1.53, respectively). Thus these would not have been selected by previous methods[15,16]. However, their joint rate of selection with GA/KNN was 4.5 times what would have been predicted based on their individual rates, suggesting that what we are detecting is their joint ability to discriminate. Another reason for selection of a gene with a t -statistic near 0 is that the GA/KNN method is inherently multivariate, and can select groups of genes that together are informative, but which marginally (one at a time) are not.

The marginal contribution of each gene in differentiating normal from tumor samples could be evaluated using the Student's t -statistic [24]. It is noteworthy that among the 50 most frequently selected genes there is no strong correlation between the magnitude of the t -statistic and the frequency of gene selection (Table 1). Although genes with a large t -statistic would be discriminative, genes with small magnitude t -statistic could also be discriminative. One reason is that samples within a class can be heterogeneous [43]. For instance, certain genes could be highly differentially expressed in one tumor subtype but not in the another. Such genes, which could well be informative, may not be identified using the t -statistic as the selection criterion, since the t -statistic would be small. In fact, the correlation between the frequency of gene selection and the absolute value of the t -statistic for the 50 most frequently selected genes is only 0.5.

For a comparison, we have applied the GA/KNN method to the leukemia data set [15]. Interestingly, for the same training set (38 specimens) used by Golub *et al.* [15], a different set of top 50 genes was selected using the GA/KNN method (multivariate) than those obtained by Golub *et al.* using the neighborhood analysis approach (univariate) [15]. Among the top 50 genes, we find that 18 and 32 genes were more highly expressed in the acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), respectively. In contrast, Golub *et al.* chose to report an equal number of more highly expressed genes in ALL and AML. When each specimen in the test set was classified according to the class memberships of its three nearest training set neighbors based on 50 genes that were selected using GA/KNN, 34 specimens were correctly classified with only one exception (AML66) (using KNN, $k = 3$ and consensus rule). Thus, the GA/KNN method correctly classified 33 of the 34 test samples. When classified using the class membership of up to 5 nearest neighbors using consensus rule (a more strict criterion), a similar result was obtained except that one specimen (AML54) became unclassifiable. Similarly, when the test set was combined with the training set, postprocessing cluster analysis of the top 50 genes using a cluster analysis program [12] showed that the AML samples and ALL samples were clustered together correctly except AML66 (Fig. 5). Furthermore, the top 50 genes found by the GA/KNN method revealed the existence of two subtypes within ALL without applying any prior knowledge. Among the 47 ALL samples, 9 were T-cell ALL (ALL2, ALL3, ALL6, ALL9, ALL10, ALL11, ALL14, ALL23, and ALL67), and the remaining

B-cell ALL. When clustered using the top 50 genes, all of the 9 T-cell ALL specimens were on one branch of the ALL tree together with two B-cell ALL (Fig. 5). This indicates that the GA/KNN method is capable of identifying genes that not only discriminate between the ALL and AML but may also unmask clinically meaningful subtypes, through subsequent cluster analysis.

In conclusion, we have described a method that selects a subset of genes that can discriminate between normal and tumor tissue samples based on microarray data. Once such a set of relevant genes has been identified, an unknown sample can be classified by comparing its expression profile with that of the known samples. In addition to this clinical application, the method could be applied to experimental settings, e.g. to characterize cellular responses to a toxic exposure. In principle, the method could be extended to toxicologic dose-response studies, to the assessment of time dependent responses, or to clinical diagnostic problems when there are more than two states. The method can be applied in a stand-alone fashion, or used as a preprocessor to cluster analysis. Our results are encouraging, with the caveat that the number of genes and the sample size used in the original expression studies are limited.

Acknowledgements

We thank Sid Soderholm, Doug Landsittel, Gene Demchuk (NIOSH/CDC), David Donson and Pierre Bushel (NIEHS) for helpful discussions. We also thank an anonymous reviewer and Dr. Uri Alon (Weizmann Institute of Science, Israel) for useful information about the misclassified specimens. We acknowledge the computational resources provided by the North Carolina Supercomputing Center. Most of the computations were carried out on SGI workstations. The work was completed when L.L. was stationed at NIEHS.

References

1. Bowtell, D.D.L. *Nature Genet.*, **1999**, 21 (suppl.), 25.
2. Brown, P.O.; Botstein, D. *Nature Genet.*, **1999**, 21 (suppl.), 33.
3. Duggan, D.J.; Bittner, M.; Chen Y.; Meltzer, P.; Trent, J.M. *Nature Genet.*, **1999**, 21 (suppl.), 10
4. Lipshutz, R.J.; Fodor, S.P.A.; Gingeras, T.R.; Lockhart D.J. *Nature Genet.*, **1999**, 21 (suppl.), 20.
5. Alizadeh, A.A.; Eisen, M.B.; Davis, R.E.; Ma, C.; Lossos, I.S.; Rosenwald, A.; Boldrick, J.C.; Sabet, H.; Tran, T.; Yu, X.; Powell, J.I.; Yang, L.; Marti, G.E.; Moore, T.; Hudson, J. Jr; Lu, L.; Lewis, D.B.; Tibshirani, R.; Sherlock, G.; Chan, W.C.; Greiner, T.C.; Weisenburger, D.D.; Armitage, J.O.; Warnke, R.; Levy, R.; Wilson, E.; Grever, M.R.; Byrd, J.C.; Botstein, D.; Brown, P.O.; Staudt, L.M. *Nature*, **2000**, 403, 503.
6. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. *Proc. Natl. Acad. Sci. USA*, **1999**, 96, 6745.
7. Perou, C.M.; Jeffrey, S.S.; van de Rijn, M.; Rees, C.A.; Eisen, M.B.; Ross, D.T.; Pergamenschikov, A.; Williams, C.F.; Zhu, S.X.; Lee, J.C.; Lashkari, D.; Shalon, D.; Brown, P.O.; Botstein, D. *Proc. Natl. Acad. Sci. USA*, **1999**, 96, 9212.
8. DeRisi, J.L.; Iyer, V.R.; Brown, P.O. *Science*, **1997**, 278, 680.
9. Ferea, T.L.; Botstein, D.; Brown, P.O.; Rosenzweig, R.F. *Proc. Natl. Acad. Sci. USA*, **1999**, 96, 9721.
10. Khan, J.; Bittner, M.L.; Saal, L.H.; Teichmann, U.; Azorsa, D.O.; Gooden, G.C.; Pavan, W.J.; Trent, J.M.; Meltzer, P.S. *Proc. Natl. Acad. Sci. USA*, **1999**, 96, 13264.
11. Voehringer, D.W.; Hirschberg, D.L.; Xiao, J.; Lu, Q.; Roederer, M.; Lock, C.B.; Herzenberg, L.A.; Steinman, L.; Herzenberg, L.A. *Proc. Natl. Acad. Sci. USA*, **2000**, 97, 2680.
12. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. *Proc. Natl. Acad. Sci. USA*, **1998**, 95, 14863.
13. Enright, A.J.; Iliopoulos, I.; Kyrpides, N.C.; Ouzounis, C.A. *Nature*, **1999**, 402, 86.
14. Marcotte, E.M.; Pellegrini, M.; Thompson, M.J.; Yeates, T.O.; Eisenberg, D. A. *Nature*, **1999**, 402, 83.
15. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; Bloomfield, C.D.; Lander, E.S. *Science*, **1999**, 286, 531.

16. Ben-Dor, A.; Bruhn, L.; Friedman, N.; Nachman, I.; Schummer, M.; Yakhini, Z. In *Proceedings of the Fourth International Conference on Computational Molecular Biology (RECOMB2000)*, ACM press: New York, **2000**.
17. Freund, Y.; Schapire, R.E. *J. Comput. Sys. Sci.*, **1997**, *55*, 119.
18. Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y.; Kaufman, L. In *Chemometrics: a textbook (Data Handling in Science and Technology, vol 2)*; Elsevier Science B. V: New York, **1988**, pp. 395-397.
19. Clark D.E.; Westhead, D.R. *J. Comput.-Aided Mol. Des.*, **1996**, *10*, 337.
20. Forrest, S. *Science*, **1993**, *261*, 872.
21. Holland, J.H. *Adaptation in Natural and Artificial Systems*, The University of Michigan Press; Ann Arbor, IL, **1975**.
22. Meza, J.C.; Judson, R.S.; Faulkner, T.R.; Treasurywala, A.M. *J. Comput. Chem.*, **1996**, *17*, 1142.
23. Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y.; Kaufman, L. In *Chemometrics: a textbook (Data Handling in Science and Technology, vol 2)*; Elsevier Science B. V: New York, **1988**, pp. 339-368.
24. Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y.; Kaufman, L. In *Chemometrics: a textbook (Data Handling in Science and Technology, vol 2)*; Elsevier Science B. V: New York, **1988**, pp. 41-48.
25. Shi, Q.; Abbruzzese, J.L.; Huang, S.; Fidler, I.J.; Xiong, Q.; Xie, K. *Clin. Cancer Res.*, **1999**, *5*, 3711.
26. Xu, L.; Xie, K.; Mukaida, N.; Matsushima, K.; Fidler, I.J. *Cancer Res.*, **1999**, *59*, 5822.
27. Wechsler, D.S.; Shelly, C.A.; Dang, C.V. *Genomics*, **1996**, *32*, 466.
28. Rasheed, B.K.; Fuller, G.N.; Friedman, A.H.; Bigner, D.D.; Bigner, S.H. *Genes Chromosomes Cancer*, **1992**, *5*, 75.
29. Gray, I.C.; Phillips, S.M.; Lee, S.J.; Neoptolemos, J.P.; Weissenbach, J.; Spurr, N.K. *Cancer Res.*, **1995**, *55*, 4800.
30. Sun, H.Q.; Yamamoto, M.; Mejillano, M.; Yin, H.L. *J. Biol. Chem.*, **1999**, *274*, 33179.
31. Asch, H.L.; Head, K.; Dong, Y.; Natoli, F.; Winston, J.S.; Connolly, J.L.; Asch, B.B. *Cancer Res.*, **1996**, *56*, 4841.
32. Afify, A.M.; Werness, B.A. *Appl. Immunohistochem.*, **1998**, *6*, 30.
33. Lee, H.K.; Driscoll, D.; Asch, H.; Asch, B.; Zhang, P.J. *Prostate*, **1999**, *40*, 14.
34. Behrens, J. *Cancer Metast. Rev.*, **1999**, *18*, 15.
35. Soler, A.P.; Knudsen, K.A.; Salazar, H.; Han, A.C.; Keshgegian, A.A. *Cancer*, **1999**, *86*, 1263.
36. Rastinejad, F.; Conboy, M.J.; Rando, T.A.; Blau, H.M. *Cell*, **1993**, *75*, 1107.
37. Hilairret, S.; Janet, T.; Pineau, N.; Caigneaux, E.; Chadeneau, C.; Muller, J.M.; Philippe, M. *Neuropeptides*, **1998**, *32*, 587.
38. Maruno, K.; Absood, A.; Said, S.I. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*, 14373.
39. Brown, M.P.; Grundy, W.N.; Lin, D.; Cristianini, N.; Sugnet, C.W.; Furey, T.S.; Ares, M. Jr.; Haussler, D. *Proc. Natl. Acad. Sci. USA*, **2000**, *97*, 262.
40. Raychaudhuri, S.; Stuart, J.M.; Altman, R.B. *Pacific Symposium on Biocomputing*, **2000**, *5*, 452.
41. Ben-Dor, A.; Shamir, R.; Yakhini, Z. *J. Comput. Biol.*, **1999**, *6*, 281.
42. Cortes, C.; Vapnik, V. *Machine Learning*, **1995**, *20*, 273.
43. Lengauer, C.; Kinzler, K.W.; Vogelstein, B. *Nature*, **1998**, *396*, 643.

Niche 1

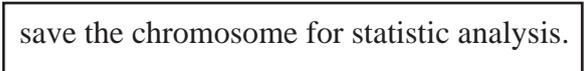
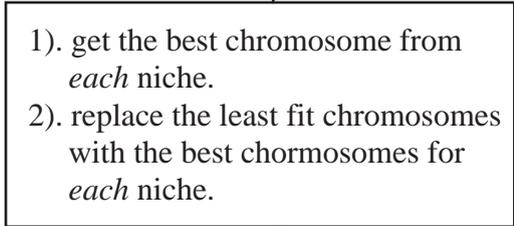
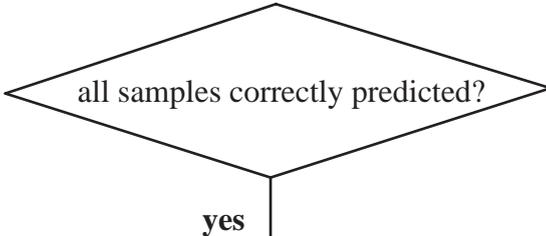
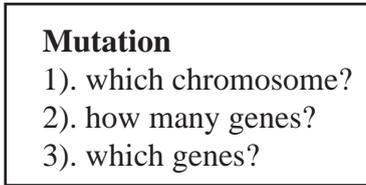
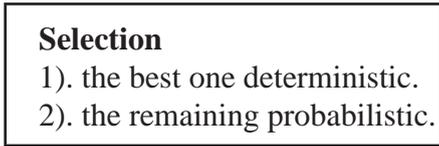
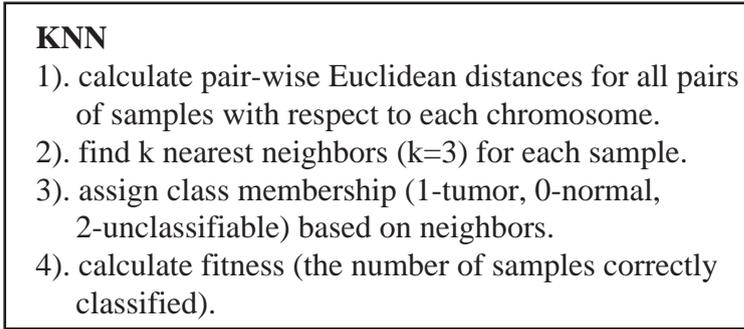
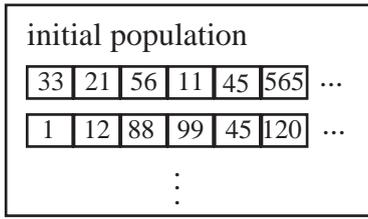


Figure 1. A schematic diagram of the GA/KNN procedure. Multiple sub-populations (niches) were performed. Only one niche is shown. At each generation, the single best chromosome found from each niche run was identified. The best chromosomes identified, one from each niche, were combined and used to replace the 10 least fit chromosomes in each niche in the next generation.

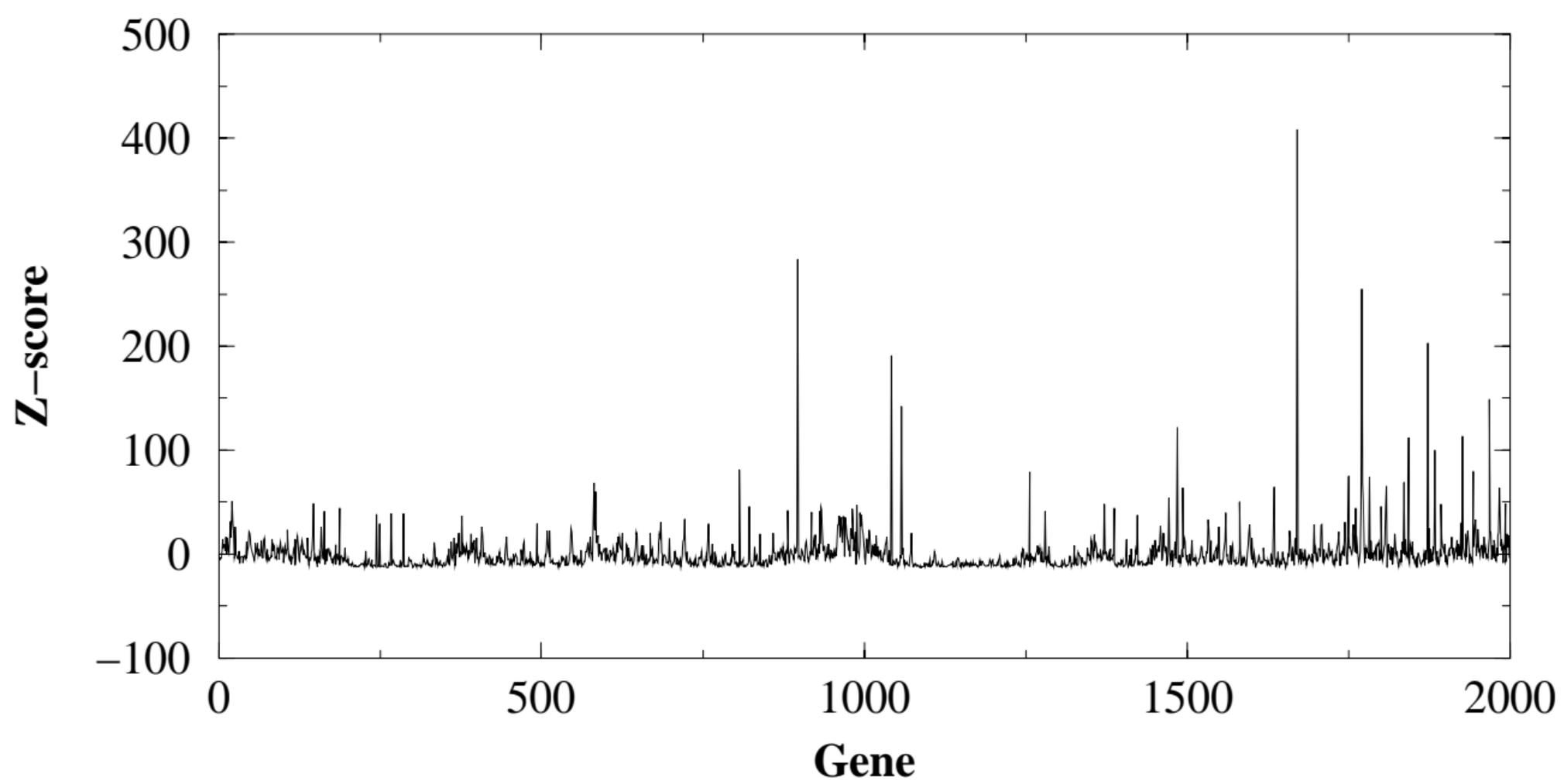
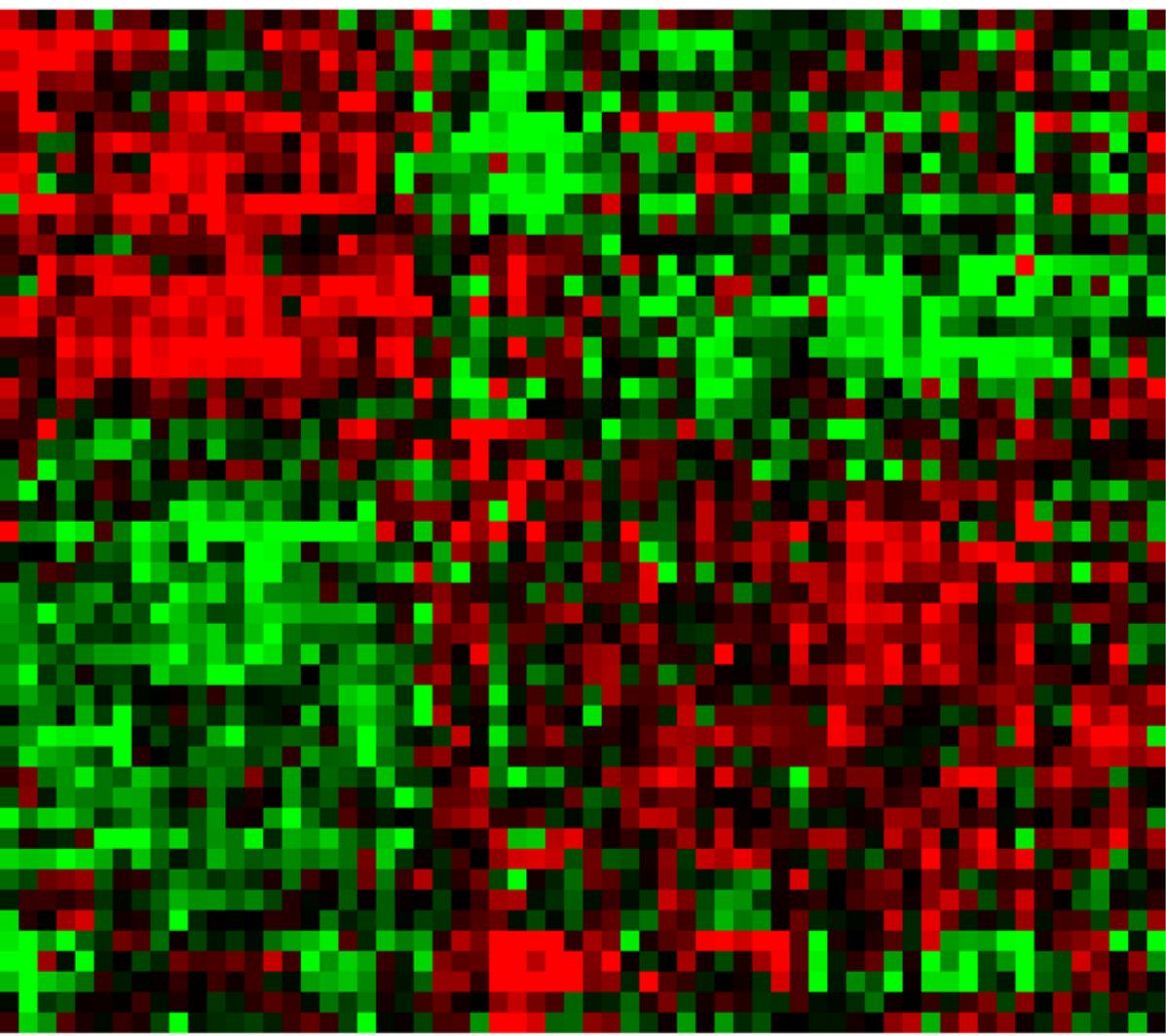
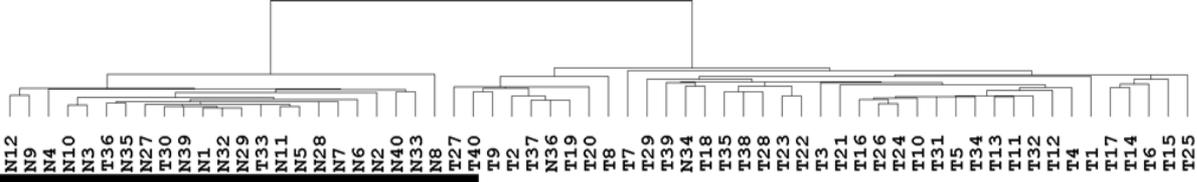
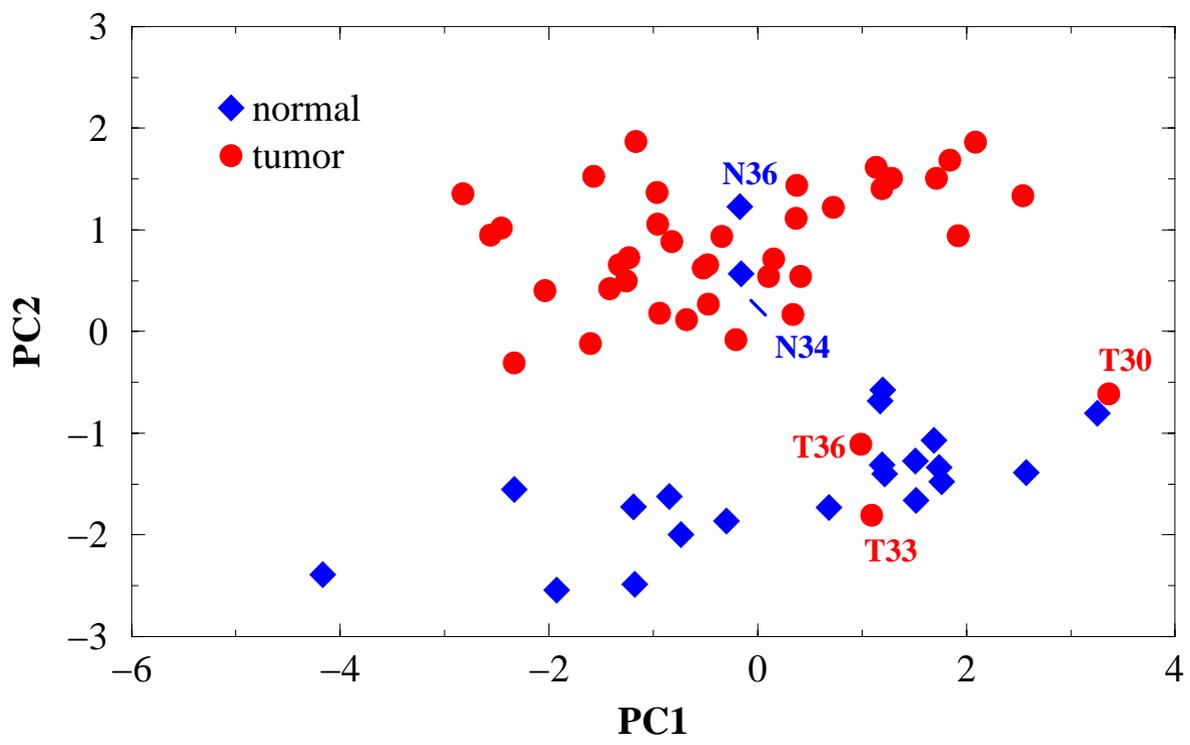
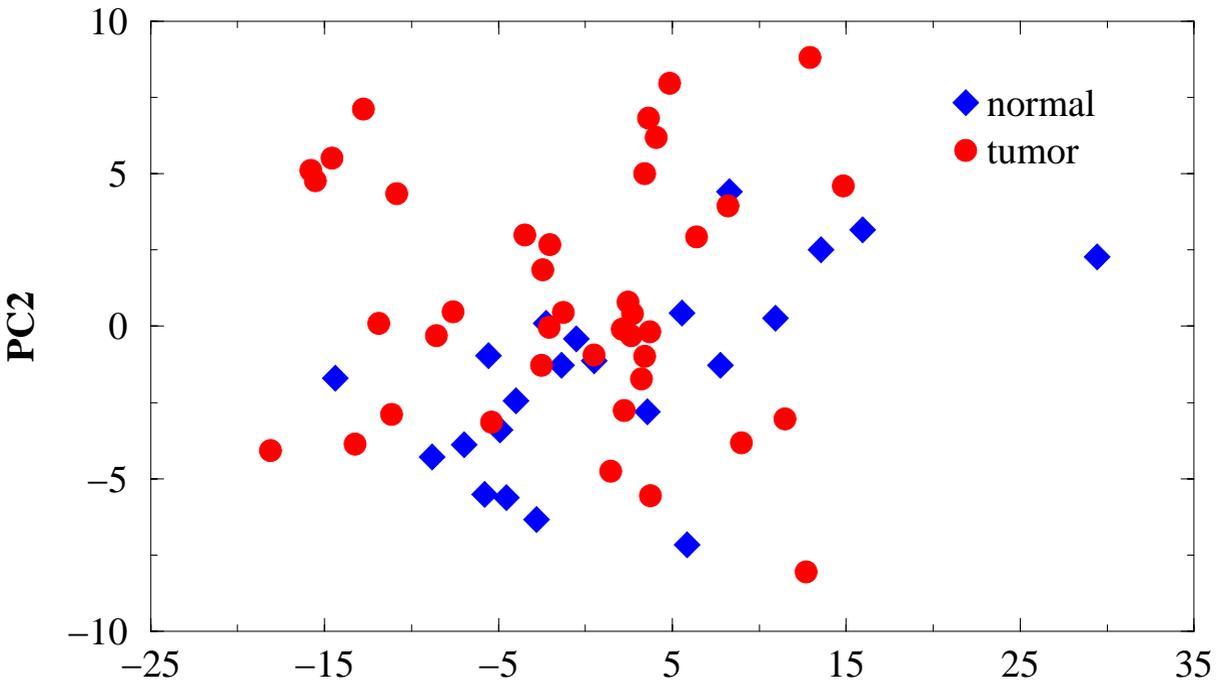


Figure 2. The statistical z -score with which each of the 2,000 genes was selected among the 6,348 solutions. Let, $Z = [S_i - E(S_i)]/\sigma$, where S_i is the number of times gene $_i$ was selected, $E(S_i)$, is the expected number of times gene $_i$ was selected, σ is the square root of the variance. Let, $A = 6,348$, $P(\text{gene}_i) = 0.025$, the probability of gene $_i$ being selected (if random). Then, $E(S_i) = P(\text{gene}_i) \cdot A$, and $\sigma = \sqrt{\{P(\text{gene}_i) \cdot [1 - P(\text{gene}_i)] \cdot A\}}$. The statistical significance $P(Z > z)$ for z -scores of 5.0 and 30.0 are approximately $3.0 \cdot 10^{-7}$ and $8.0 \cdot 10^{-198}$, respectively.



METALLOTHIONEIN-II
 TALLA-1
 H.sapiens a-L-fucosidase
 KERATIN, TYPE II CYTOSKELETAL 8
 Human MXI1
 Human mucin 2 (MUC2)
 11 beta-hydroxysteroid dehydrogenase
 transmembrane carcinoembryonic antigen BGPC
 transmembrane carcinoembryonic antigen BGPa
 PHOSPHOENOLPYRUVATE CARBOXYKINASE
 MINERALOCORTICOID RECEPTOR
 cAMP-dependent protein kinase catalytic
 PERIPHERAL MYELIN PROTEIN 22
 TROPOMYOSIN
 COMPLEMENT FACTOR D
 vasoactive intestinal peptide (VIP)
 H.sapiens mRNA for hevin like protein
 GELSOLIN PRECURSOR
 alcohol dehydrogenase class I
 LCA-homolog. LAR protein
 HLA CLASS II HISTOCOMPATIBILITY ANTIGEN
 60S ACIDIC RIBOSOMAL PROTEIN P2
 PROBABLE G PROTEIN-COUPLED RECEPTOR 6H1
 INTERFERON-INDUCIBLE PROTEIN 1-8D
 hormone-sensitive lipase (LIPE)
 MONAP
 INTERFERON-INDUCIBLE PROTEIN 9-27
 DNA polymerase delta small subunit
 EUKARYOTIC INITIATION FACTOR 4B
 Human aspartyl-tRNA synthetase alpha-2
 HEAT SHOCK PROTEIN HSP 90-BETA
 TRANSCRIPTION FACTOR IIIA
 COLLAGEN ALPHA 2 (XI)
 MEMBRANE COFACTOR PROTEIN PRECURSOR
 RIBOPHORIN II PRECURSOR (HUMAN)
 chloride channel regulatory protein
 Human 100 kDa coactivator
 HEK2 protein tyrosine kinase receptor
 H.sapiens mRNA for beta-COP
 pterin-4a-carbinolamine dehydratase
 p cadherin
 CALGIZZARIN
 tyrosine kinase receptor p145TRK-B
 X BOX BINDING PROTEIN-1 (HUMAN)
 GLUCOCORTICOID RECEPTOR, BETA
 OSF-2p1
 pro-alpha1(III) collagen
 THROMBOSPONDIN 2 PRECURSOR
 Human TGF-beta type II receptor
 YRRM

Figure 3. Hierarchical clustering [12] of gene expression data for the colon data using the top 50 genes.



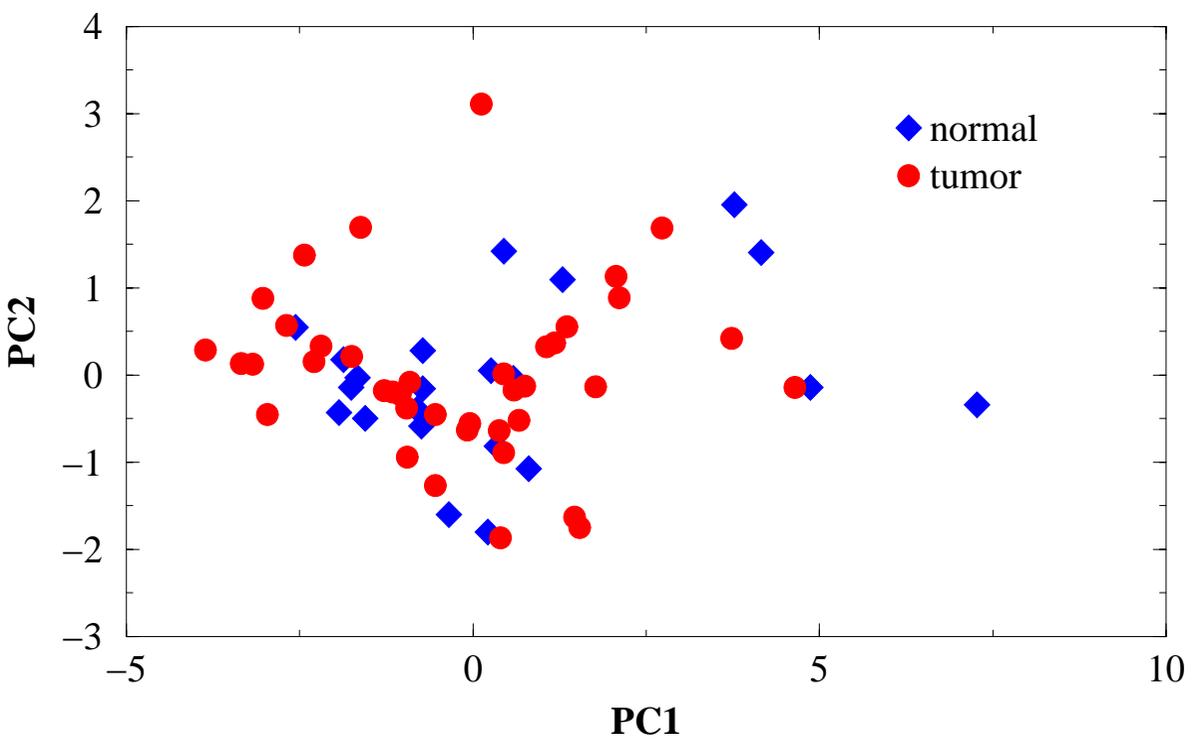
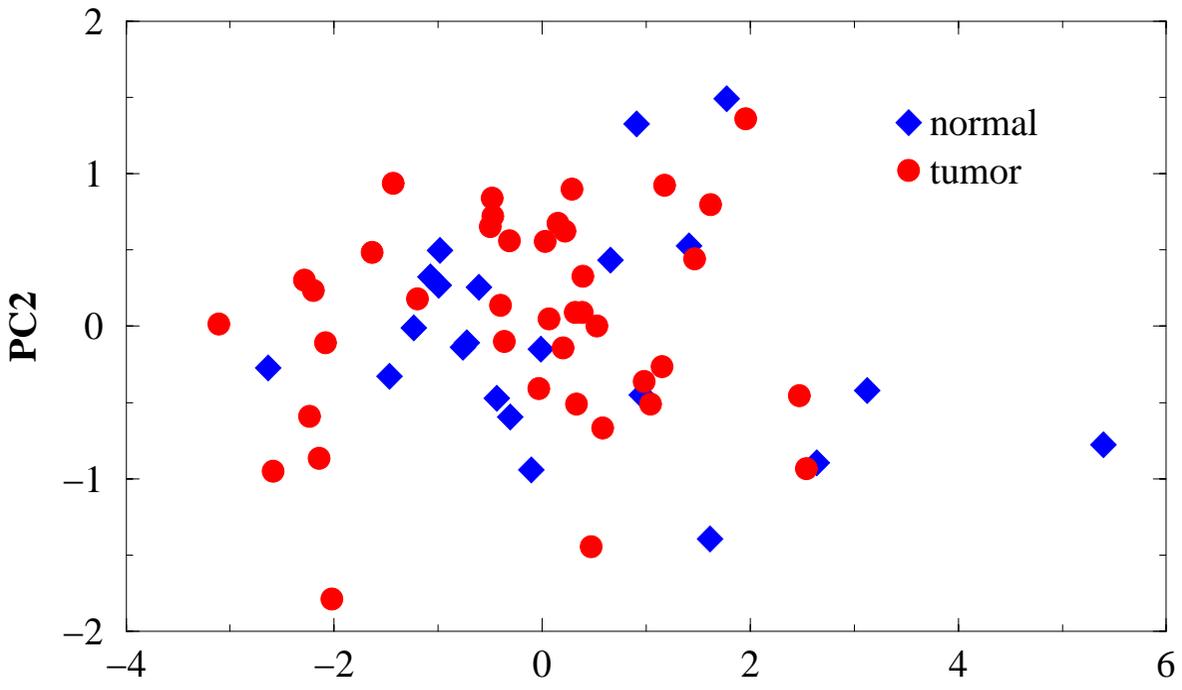
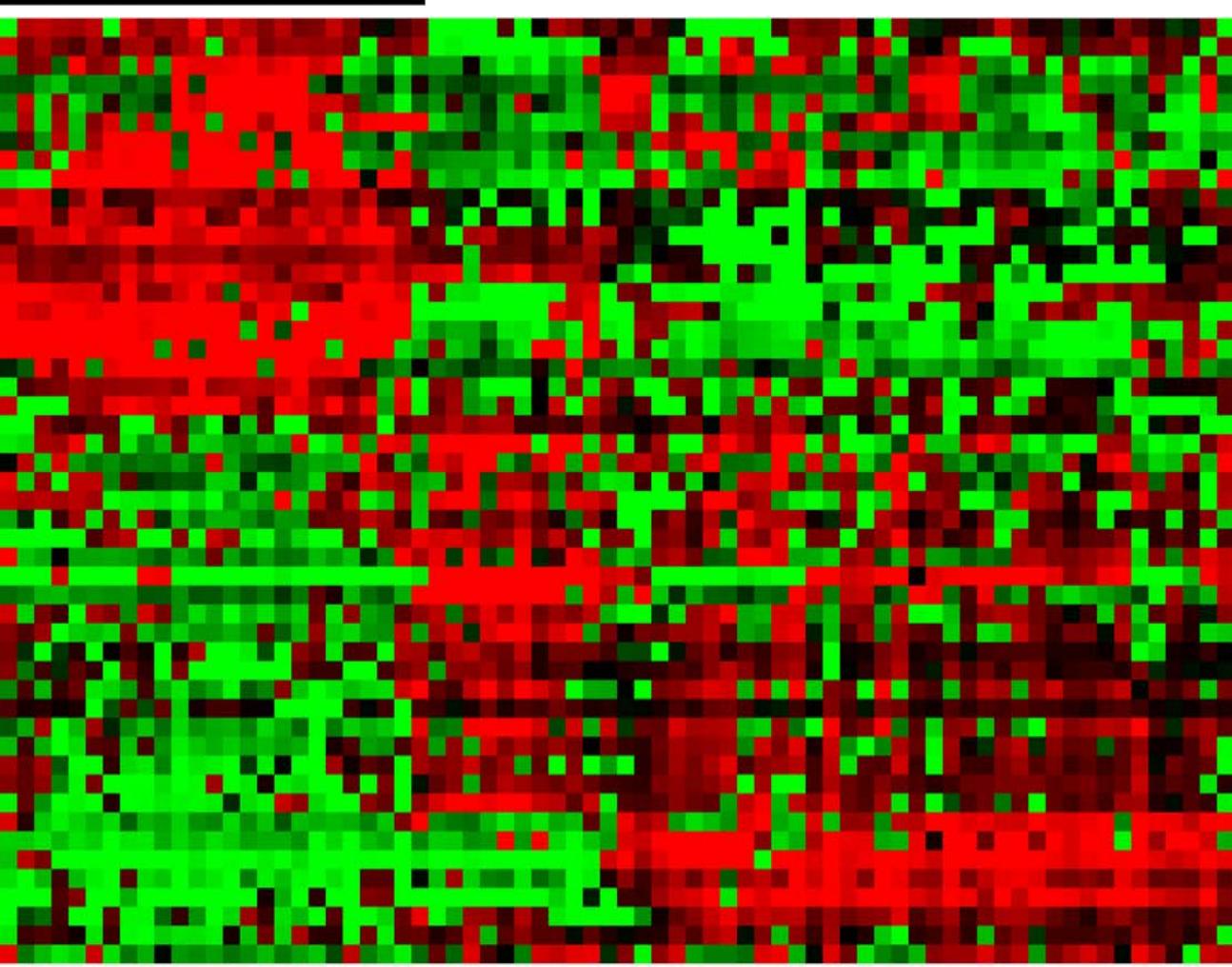
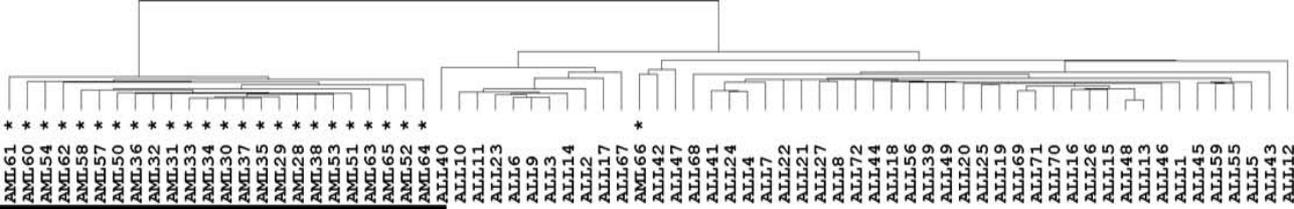


Figure 4. Plot of the first versus the second principal component. *a*, using the expression levels (\log_{10} transformed) of all 2,000 genes. *b*, using the 50 most frequently selected genes. *c*, using the 50 randomly selected genes. *d*, using the 100 least frequently selected genes.



GLUL Glutamate-ammonia ligase
 HMOX2 Heme oxygenase (decycling) 2
 Ahnak-A Nucleoprotein Ahnak-A
 Erythrocyte membrane 50kd glycoprotein
 CA2 Carbonic anhydrase II
 NFIX Nuclear factor I/X
 HOMEBOX PROTEIN HOX-A5
 Homeotic Protein C6, Class I
 KIAA0246 gene, partial cds
 ELA2 Elastatse 2, neutrophil
 MPO Myeloperoxidase
 CST3 Cystatin C
 ARHG Ras homolog, member G (rho G)
 ANX1 Annexin I (lipocortin I)
 CSF3R
 SPI1 integration oncogene
 D component of complement
 NF-IL6-beta
 Ahnak-Related Sequence
 MANA2 Alpha mannosidase II isozyme
 GRO2 oncogene
 Homolog suppressor-of-white-apricot
 CD2 antigen (p50)
 Myosin VIIA (USH1B)
 SPRR1B Small proline-rich protein 1B
 ACTN2 Actinin alpha 2
 KIAA0080 gene, partial cds
 DTYMK Deoxythymidylate kinase
 Carcinoembryonic antigen precursor
 LTB Lymphotoxin-beta
 GATA3 GATA-binding protein 3
 GB DEF = Escherichia coli unknown
 GUANYLATE CYCLASE, BETA-1
 Spinal Muscular Atrophy 4
 KIAA0239 gene, partial cds
 Butyrophilin (BTF4)
 MHC-encoded proteasome LAMP7-E1
 RPA1 Replication protein A1
 Clone 22 mRNA
 CTPS CTP synthetase
 UBIQUITIN-LIKE PROTEIN GDX
 (AF1q) mRNA
 S100 calcium-binding protein A13
 BLK Protein-tyrosine kinase blk
 IGB (B29)
 OBF-1
 TCL1
 CD19 antigen
 Skeletal muscle abundant protein
 Clone 23612 mRNA sequence

Figure 5. Hierarchical clustering [12] of gene expression data for the leukemia data using the top 50 genes. The original data were downloaded from the web (<http://www.genome.wi.mit.edu/MPR>). The data set contained a training set (38 specimens) and a test set (34 specimens). The data set were *log10* transformed. Only the training set samples were used to obtain 10,000 subsets of 50 genes that discriminate the acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). A subset of 50 genes was considered discriminative when 37 of the 38 specimens in the training set were correctly classified by the GA/KNN. The frequency of gene selection was subsequently analyzed. The 50 most frequently selected genes were used to cluster all the specimens (training and test) and are showed here. It can be seen that all ALL and AML specimens were clustered together separately with only one exception (AML66). Thus, the GA/KNN method correctly classified 33 of the 34 test samples. Furthermore, it seems that the top 50 genes obtained by the GA/KNN method revealed the two subtypes within ALL without applying any prior knowledge. Among the 47 ALL samples, 9 (ALL2, ALL3, ALL6, ALL9, ALL10, ALL11, ALL14, ALL23, and ALL67) were T-cell ALL and the remaining B-cell ALL. It appears that all of the 9 T-cell ALL samples were on one branch of the ALL tree together with two B-cell ALL.

Table 1 • The 50 most frequently selected genes^a

Gene Number	Student's <i>t</i> -statistic ^b	Name
M26383	7.21	Human monocyte-derived neutrophil-activating protein (MONAP)
H43887	-5.08	Complement factor D precursor
J05032	6.77	Human aspartyl-tRNA synthetase alpha-2 subunit
L07648	-1.96	Human MXI1
R36977	5.46	P03001 transcription factor IIIA
L12350	2.62	Thrombospondin 2 precursor
M80815	-5.40	<i>H. sapiens</i> a-L-fucosidase gene, exon 7 and 8
D13665	2.91	Human mRNA for osteoblast specific factor 2 (OSF-2p1)
U36621	1.40	Human Y-chromosome RNA recognition motif protein (YRRM)
H06524	-3.16	Gelsolin precursor, plasma
R44301	-3.89	Mineralocorticoid receptor
M94132	-3.01	Human mucin 2 (MUC2)
D29808	-4.18	Human T-cell acute lymphoblastic leukemia associated antigen 1
U22055	4.04	Human 100 kDa coactivator
T54303	-2.12	Keratin, type II cytoskeletal 8
X06700	2.33	Human 3' region for pro-alpha1(III) collagen
U14631	-2.58	Human 11 beta-hydroxysteroid dehydrogenase type II
T51571	4.79	P24480 calgizzarin
H08393	4.71	Collagen alpha 2(XI) chain
H24310	1.38	Glucocorticoid receptor, beta
M36634	-4.65	Human vasoactive intestinal peptide (VIP)
U12140	0.44	Human tyrosine kinase receptor p145TRK-B (TRK-B)
X86693	-3.75	<i>H. sapiens</i> mRNA for hevin like protein
M31627	1.83	X box binding protein-1
L41559	4.37	<i>H. sapiens</i> pterin-4a-carbinolamine dehydratase (PCBD)

R75893	1.20	Probable G protein-coupled receptor 6H1 from T-cells
H79852	1.53	60S acid ribosomal protein P2
X63629	4.74	<i>H. sapiens</i> mRNA for p cadherin
R34698	4.36	Interferon-inducible protein 9-27
L11706	2.97	Human hormone-sensitive lipase (LIPE) gene
X75208	3.06	<i>H. sapiens</i> HEK2 mRNA for protein tyrosine kinase receptor
X07767	-1.38	Human cAMP-dependent protein kinase catalytic subunit type α
T90280	3.78	Ribophorin II precursor
H26965	0.18	HLA class II histocompatibility antigen, gamma chain precursor
T92451	-3.96	Tropomyosin, fibroblast and epithelial muscle-type
M85079	2.00	Human TGF-beta type II receptor
T94350	-2.67	Peripheral myelin protein 22
T51023	5.28	Heat shock protein HSP 90-beta
L05144	-3.22	Phosphoenolpyruvate carboxykinase, cytosolic
X82103	2.59	<i>H. sapiens</i> mRNA for beta-COP
U17899	4.16	Human chloride channel regulatory protein
U21090	4.30	Human DNA polymerase delta small subunit
Y00815	-0.01	Human LCA-homolog. LAR protein (leukocyte antigen related)
R33367	3.78	Membrane cofactor protein precursor
M12272	-1.08	<i>H. sapiens</i> alcohol dehydrogenase class I γ subunit (ADH3)
X57351	3.75	Interferon-inducible protein 1-8D
X16354	-2.35	transmembrane carcinoembryonic antigen BGP α (formerly TM1-CEA)
X16356	-3.21	transmembrane carcinoembryonic antigen BGPC (formerly TM3-CEA)
T51858	3.69	Eukaryotic initiation factor 4B
R06601	-1.43	Metallothionein-II

^aThe genes are listed in descending order based on the rank frequency obtained using the training set samples (see text for details). A complete list of the 2,000 genes is available on <http://dir.niehs.nih.gov/microarray/datamining/>.

^bStudent's two-sample *t*-test was performed on \log_{10} transformed data of the 42 training set samples. A positive value indicates that the gene is up regulated in tumors compared to normal samples and conversely. The *t*-values at which *P* is 0.1, 0.01, and 0.001 are 1.684, 2.704, and 3.551, respectively.

Table 2 • Classification of the test set samples^a

Sample	Exp. ^b	Top 1 ^c	Top 5 ^c	Top 25 ^c	Top 50 ^c	Top 100 ^c	Top 500 ^c	All 2000	The least 100 ^c
N29	0	1	0	0	0	0	2	2	2
N32	0	0	0	0	0	0	0	2	2
N33	0	1	1	0	0	0	0	0	2
N34	0	1	1	1	1	1	1	1	2
N35	0	1	2	0	0	0	0	2	2
N36	0	1	1	1	1	1	1	1	2
N39	0	1	0	0	0	0	0	0	2
N40	0	1	1	0	0	0	0	2	2
T29	1	1	1	1	1	1	1	1	2
T30	1	2	2	0	0	0	2	1	1
T31	1	1	1	1	1	1	1	1	1
T32	1	1	1	1	1	1	1	1	1
T33	1	2	0	0	0	0	2	2	2
T34	1	1	1	1	1	1	1	2	2
T35	1	1	1	1	1	1	1	1	1
T36	1	1	1	0	0	0	0	2	2
T37	1	1	1	1	1	1	2	2	2
T38	1	1	1	1	1	1	1	1	1
T39	1	1	1	1	1	1	1	1	2
T40	1	1	1	1	1	1	1	1	2

^aA sample is classified as 0-normal, 1-tumor, or 2-unclassifiable. See text for details.

^bFrom ref 6.

^cClassification using the top most and bottom least frequently selected genes.