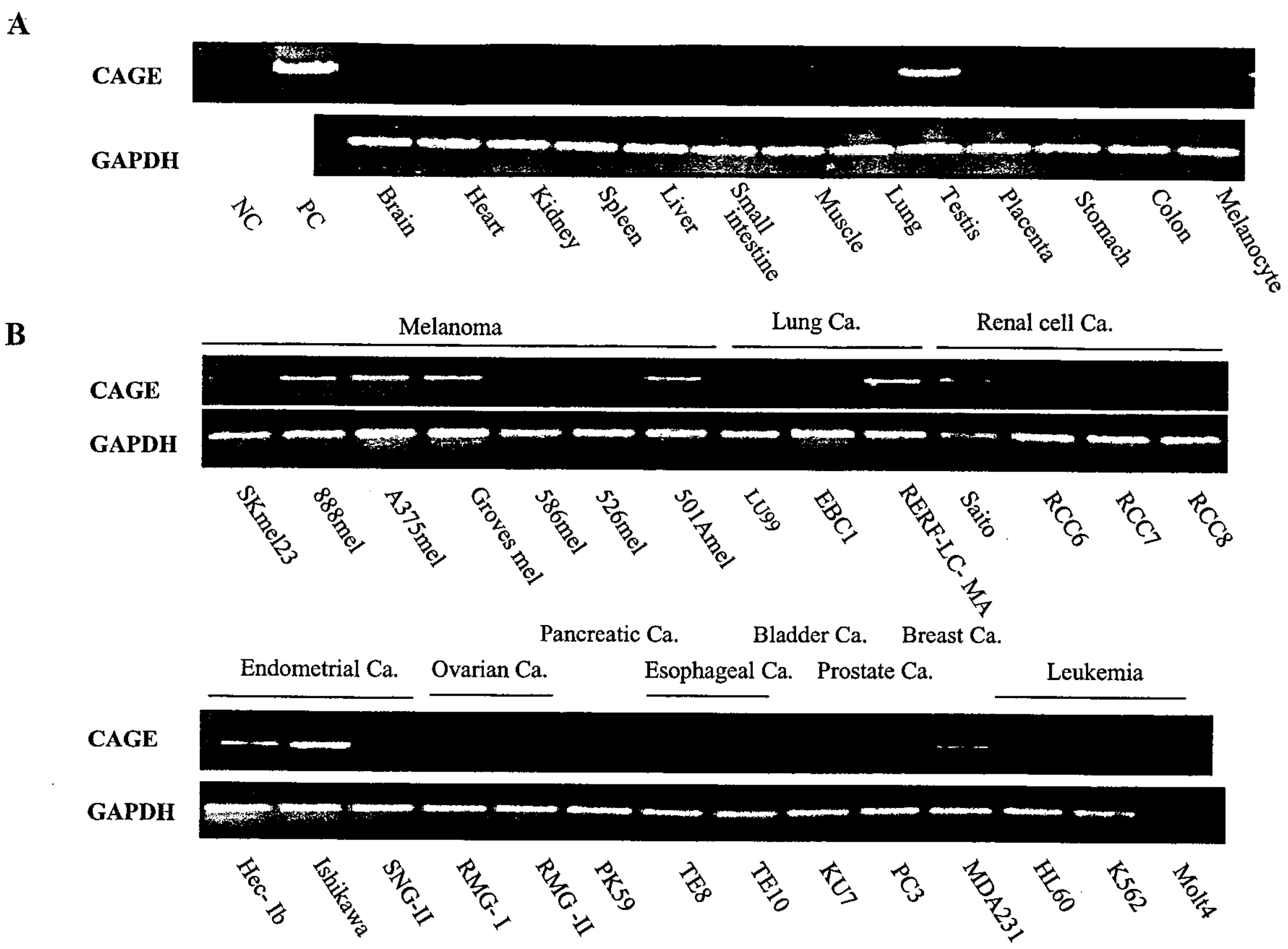
The specific CAGE band was detected in lysates from 2 endometrial cancer cell lines, Hec-Ib and Ishikawa, and one melanoma cell line 888mel, those are CAGE positive by RT-PCR analysis, but was not shown in lysate from PCR negative cultured melanocytes. NIH-3T3 cells transfected with pcDNA-CAGE was positive control and untransfected NIH-3T3 cells was negatine control.

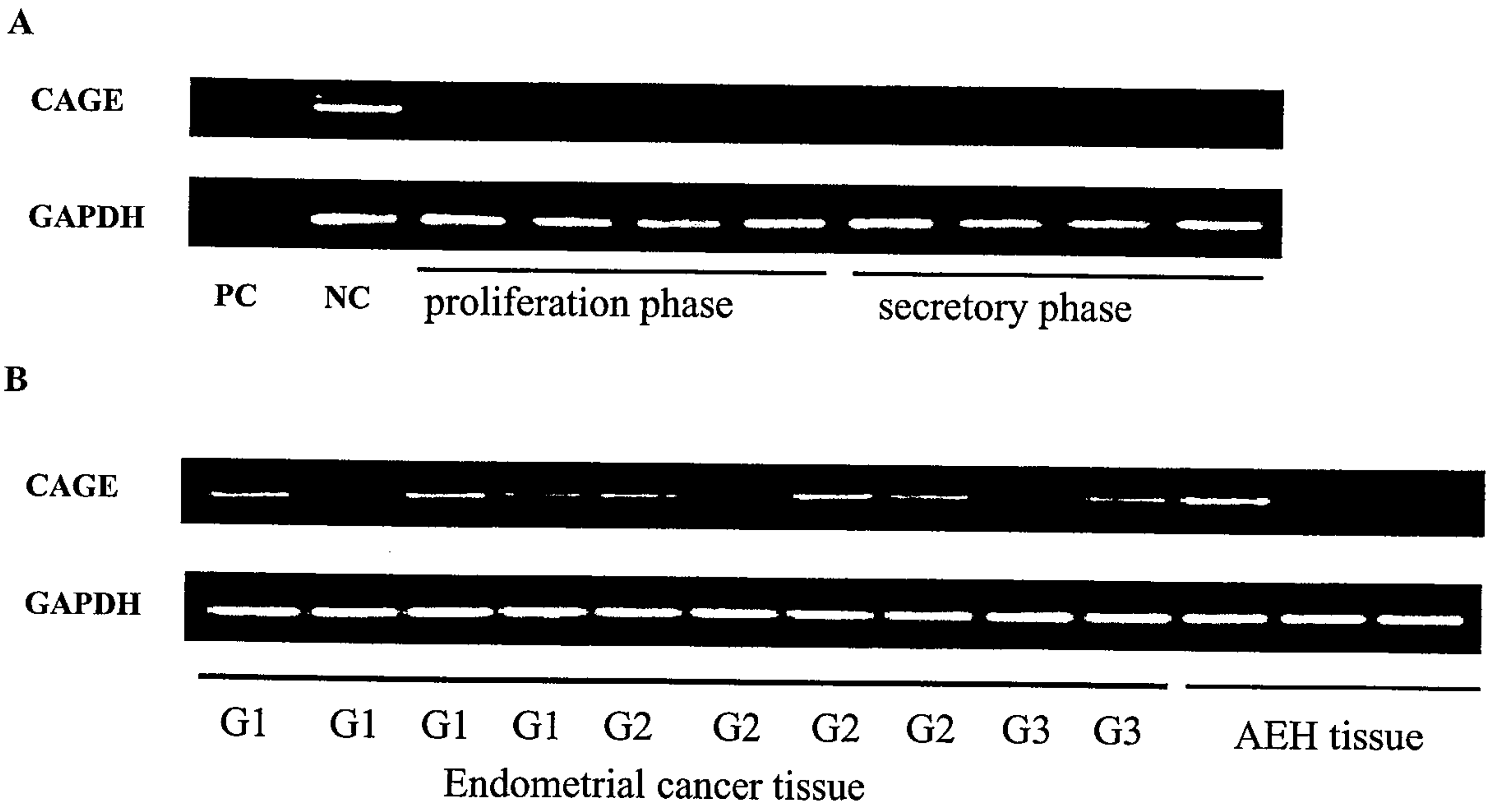
Fig. 4. Presence of anti-CAGE IgG antibodies in sera from various cancer patients detected by Western blot analysis with bacterial recombinant CAGE protein

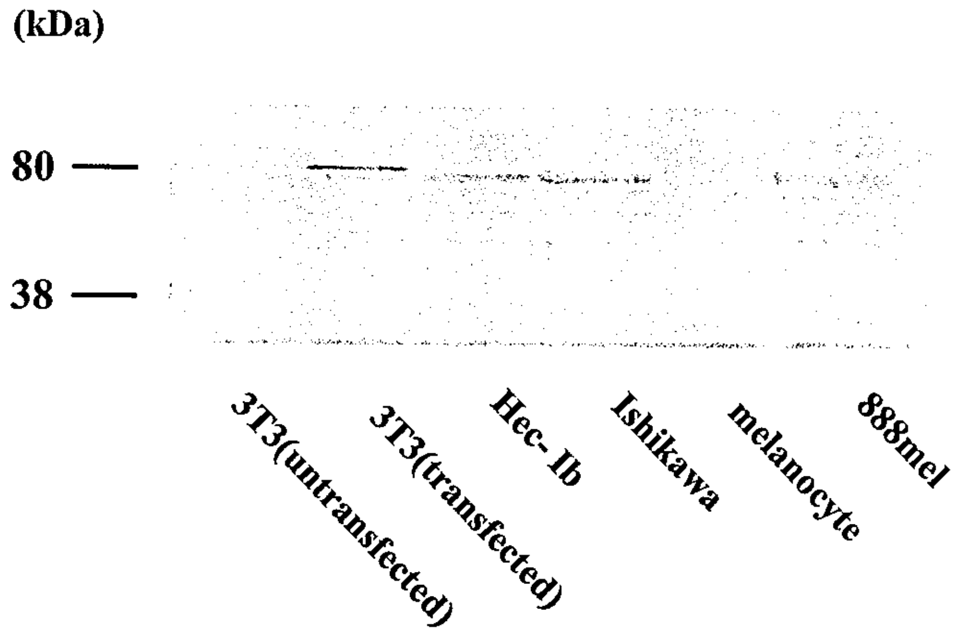
By Western blot analysis, the recombinant His-tagged CAGE protein fragment containing N-terminal 211 amino acids of CAGE (M.W. = 31.1kDa) was recognized by IgG antibodies in sera from some of the patients with various cancers. Lane1: Staining of CAGE with anti-His antibody. Lanes2-12: staining with 1:100 diluted sera. Lanes 2-6 sera from endometrial cancer patients; Lanes 7and 8 sera from melanoma patients; Lane 9 and 10 sera from colon cancer patients; Lane 11 and 1; sera from healthy controls. Only positive cancer samples are shown in this representative experiment. No band was shown in the lanes with sera from 2 healthy individuals. 1μg of recombinant CAGE protein was loaded per lane.

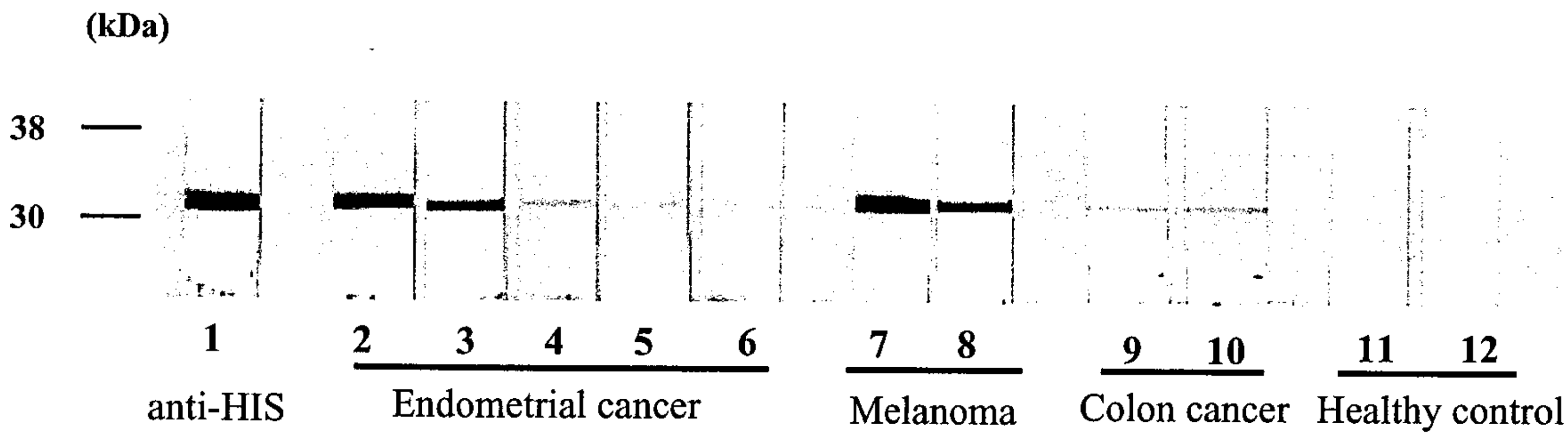
Fig.5 Frequent detection of anti-CAGE antibodies in sera from patients with MSI-H endometrial cancer evaluated by ELISA

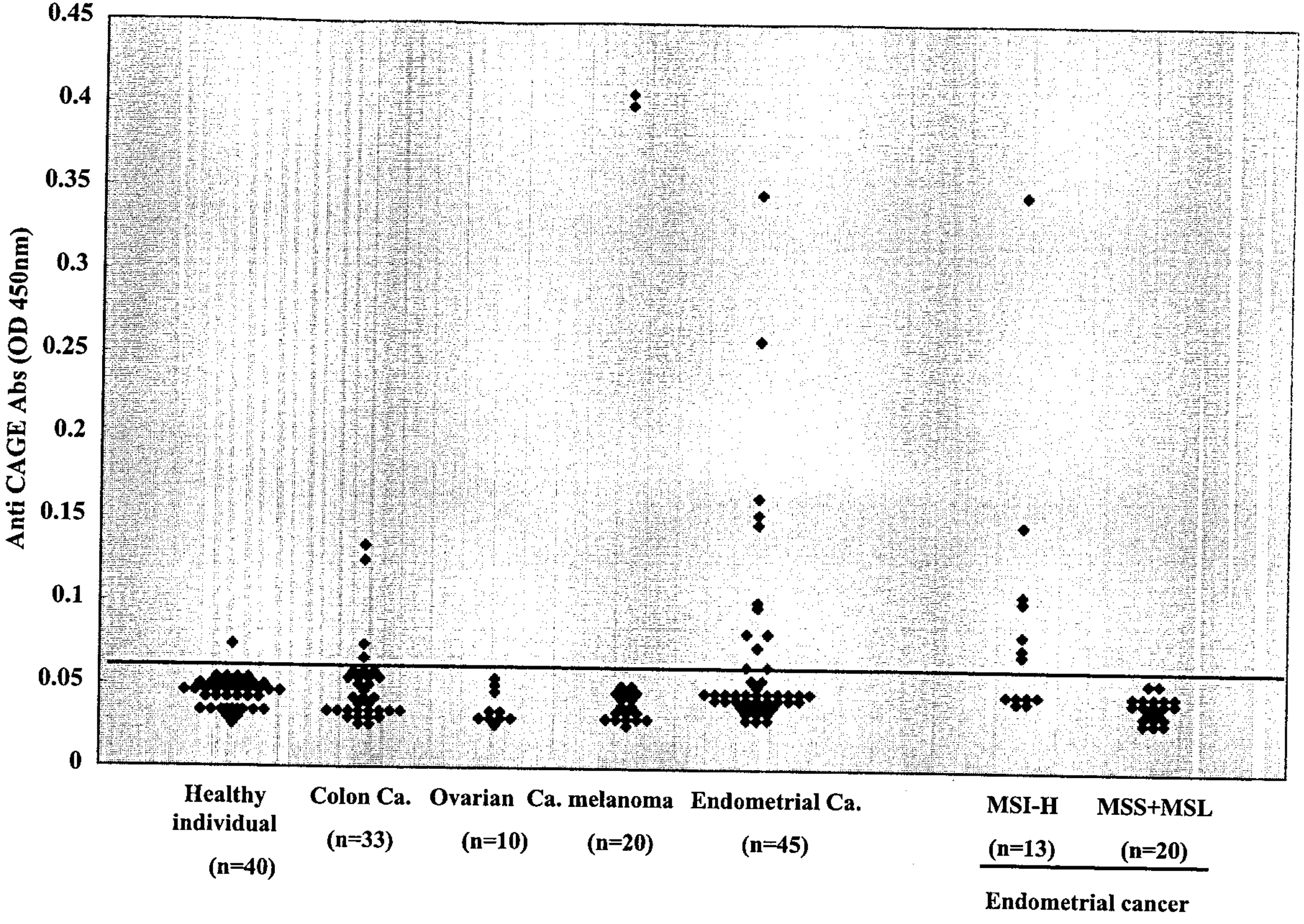
ELISA was performed with the recombinant CAGE protein. The horizontal line indicates the cutoff value for positivity (OD=0.058: the average absorbance of the healthy individuals plus 2SD). Positive sera were found in 12 of 45 (26.7%) endometrial, 4 of 33 (12.5%) colon cancer, 2 of 20 (10.0%) melanoma patients, and 1 of 40 (2.5%) age matched healthy individuals, but not in 10 ovarian cancer patients. Among 33 endometrial cancer patients whose MSI status was evaluated, 7 of 13 (53.8%) patients with MSI-H had positive serum CAGE antibody, while none of 20 non-MSI-H patients including one MSI-L and 19 MSS patients had anti-CAGE antibody.


A
CAGE
GAPDH
NC
PC
Brain
Heart
Kidney
Spleen
Liver
Small intestine
Muscle
Lung
Testis
Placenta
Stomach
Colon
Melanocyte
B
Melanoma
Lung Ca.
Renal cell Ca.
CAGE
GAPDH
SKmel23
888mel
A375mel
Groves mel
586mel
526mel
501Amel
LU99
EBC1
RERF-LC-MA
Saito
RCC6
RCC7
RCC8
Pancreatic Ca.
Bladder Ca.
Breast Ca.
Endometrial Ca.
Ovarian Ca.
Esophageal Ca.
Prostate Ca.
Leukemia
CAGE
GAPDH
Hec- Ib
Ishikawa
SNG-II
RMG-I
RMG-II
PK59
TE8
TE10
KU7
PC3
MDA231
HL60
K562
Molt4


A
CAGE
GAPDH
PC
NC
proliferation phase
secretory phase
B
CAGE
GAPDH
G1
G1
G1
G1
G2
G2
G2
G2
G3
G3
Endometrial cancer tissue
AEH tissue


(kDa)
80
38
3T3(untransfected)
3T3(transfected)
Hec- Ib
Ishikawa
melanocyte
888mel


(kDa)
38
30
1
2
3
4
5
6
7
8
9
10
11
12
anti-HIS
Endometrial cancer
Melanoma
Colon cancer
Healthy control


Anti CAGE Abs (OD 450nm)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Healthy individual (n=40)
Colon Ca. (n=33)
Ovarian Ca. (n=10)
melanoma (n=20)
Endometrial Ca. (n=45)
MSI-H (n=13)
MSS+MSL (n=20)
Endometrial cancer

all existing patterns of expression that could be useful in classification. This is made possible by introducing clustering analysis techniques, such as K-means clustering, to feature selection.

For the classification of tumors, many machine learning algorithms are available. Besides the simple methods like weighted voting scheme [1] and K nearest neighbor(KNN), support vector machine (SVM) has been widely used by many researchers. Khan *et al.* demonstrated the application of artificial neural networks for discriminating four subtypes of the small, round blue cell tumors (SRBCTs) of childhood[5]. Nevertheless, some comparative studies seem to suggest that simple algorithms tend to have a higher reliability than more complicated ones[11].

In the choice of classification algorithms, we find it important to ask the following questions. If a classifier is trained to discriminate, for example, two subtypes of leukemia, what kind of prediction will it produce for a sample of a newly discovered subtype it has never seen? What if the classifier is presented with normal tissues, or even tissues of stomach cancer? Ideally, these samples should not be classified as either of the two subtypes; otherwise, it would be counted as false positive. Therefore, the above questions lead to the test of false positives. Validation of classifiers in previous studies has been mainly focusing on the false negative cases as most samples for independent tests belong to one of the training subtypes. Despite the importance of avoiding false positives, especially in the process of defining new cancer subtypes and in the detection of metastatic cancers, extensive test of false-positive error rates of various classification schemes have not been reported in the literature.

In this paper, we test the false positive rates of various classification schemes through a 'null-test' in which a classifier is presented with a large number of samples that do not belong to any of the tumor types in the training dataset. To achieve a relatively large dataset, data from 239 microarray experiments performed in several laboratories are pooled together to test the false positive of one classifier. We compare both the false-positive and false-negative error rate of KNN, SVM, and prototype matching (PM), which is perhaps the simplest pattern recognition technique.

2 Method

The whole process of our approach is summarized in Fig. 1A.

2.1 Statistical feature selection

Kruskal-Wallis H test. Kruskal-Wallis H test is the non-parametric counterpart of analysis of variance (ANOVA), which is a standard statistical tool for detecting differences in multi-group comparison. We choose the non-parametric test because it avoids making the assumption that the expression levels are normally distributed with equal variances within groups. It is believed that nonparametric statistical tests are nearly as powerful in detecting differences among populations as parametric methods when the data are normal. They are more powerful in situations where the data does not meet the underlying assumptions of parametric methods. Some statisticians advocate the use of nonparametric methods.

For each gene, a statistic H is calculated according to the ranks of its expression levels between multiple groups. The score is defined as: $H = \frac{12}{N(N+1)} \sum \frac{r_i^2}{n_i} - 3(N+1)$, where N is the number of tumor types in question, r_i is sum of ranks of tumor type i which has n_i samples. The higher H is, the higher the degree of association. The score tells us to what extent the gene expresses differently between *any* two groups. This score is directly related to P values because it follows a χ^2 distribution with $N-1$ degrees of freedom. For $N = 3$ case, if a gene's score is above the critical value of 9.21, we can tell with $P<0.01$ that this gene correlates significantly with group distinctions. Genes that fail to reach this level of significance are eliminated without further analysis. Note that the statistical significance does not decay with the increase of N. The reason statisticians invent ANOVA and Kruskal-Wallis H test is to avoid the accumulation of errors in multiple pairwise T-test or Mann-Whiteny U test. This

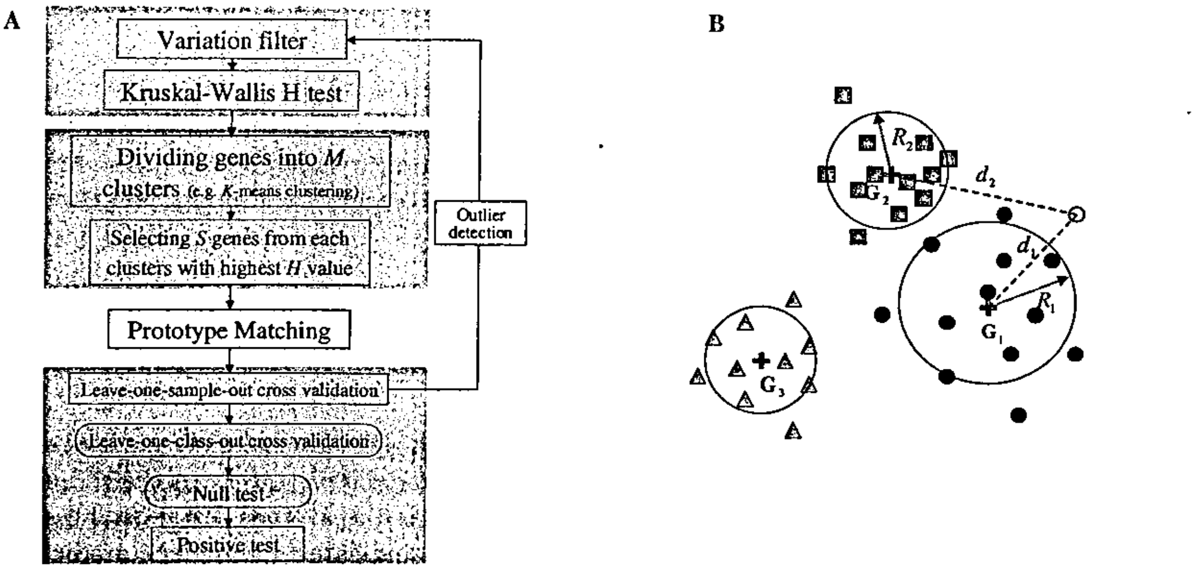

A
Variation filter
Kruskal-Wallis H test
Dividing genes into M clusters (e.g. K-means clustering)
Selecting S genes from each clusters with highest H value
Outlier detection
Prototype Matching
Leave-one-sample-out cross validation
Leave-one-class-out cross validation
Null test
Positive test
B
R_2
G_2
d_2
d_1
R_1
G_1
G_3

Figure 1: **A,** Outline of cancer classification procedure. Based on non-parametric statistics, a cluster-and-select strategy is employed in the selection of informative genes. False positive errors are tested by null test and leave-one-class-out cross validation (LOCOCV). **B,** Prototype matching. A new sample (open circle) is compared with the existing prototypes.

could happen in *one-vs-all* and *all-vs-all* approaches, where the selection of informative genes is based on $O(N)$ and $O(N^2)$ statistical tests, respectively.

Redundancy reduction: classification of genes for the classification of samples. All those genes that passed the H test convey information that could be useful in classification. But still there are too many of them. We noted that many genes have very similar expression patterns. So it is possible to reduce the size of feature set without incurring classification accuracy. This is the so-called redundancy reduction problem in feature selection[12].

Another issue is that the H score does not tell us which pair-wise distinction a certain gene is associated with. It is possible that there are more genes associated with A-B distinction than those with B-C and C-A, when three subtypes A, B, and C are considered. To improve the overall accuracy of classification, the choice of genes should be made in balance.

We tackle these two problems at the same time through a cluster-and-select strategy. The idea is to select a relatively small number of representatives from each cluster of similarly expressed genes. Methods for clustering analysis have been the subject of extensive research in bioinformatics, and there exist many algorithms. Here we borrowed such techniques for another purpose: gene filtering. We used the simple k-means clustering method, which divides a set of genes into a predefined number of clusters by maximizing the between group variance. In the resultant grouping, some clusters may contain more genes than others. But we select a given number (S) of genes from all clusters to increase the signal-to-noise ratio in classification. Because the H score indicates the significance of association, the genes with higher scores are selected from each group. This is what we call a local filter.

2.2 Prototype matching

Prototype matching is a simple method for pattern recognition. Basically, it stores prototypes and compares a new sample with them. As depicted in Fig. 1B, each tumor type is characterized by an expression prototype G_k and a radius of cluster R_k, where $k = 1, 2, 3$. Prototypes are simply calculated as the average of the expression pattern in training samples, while the radius of each cluster is the average distance between samples and this prototype. Distances are calculated by: $d_{ij} = 1 - P_{ij}$, where

P_{ij} is the Pearson's correlation coefficients between two expression patterns i and j. As $P_{ij} \in [-1,1]$, we have $d_{ij} \in [0,2]$.

To demonstrate how PM algorithm works, we use the configuration in Fig. 1B as an example. For a new sample shown in the top right, we calculate its distance to all prototypes and find that d_1, the distance to prototype G_1, is the shortest. Therefore, it is temporarily assigned to type 1. The distance to the second nearest prototype G_2 is also calculated (d_2). The confidence of prediction is measured by the following scores:

$$\begin{aligned} m &= (d_2 - d_1)/d_2, & (1)\\ d_r &= d_1/R_1, & (2)\\ C &= m/d_r. & (3)\end{aligned}$$

The parameter m characterize the margin of the winner prototype. For an ideal match, where $d_1 \ll d_2$, we have $m \approx 1$. By calculating the parameter d_r, we compare the distance d_1 with the radius of the prototype G_1. Ideally d_r should be about 1.0 or smaller, as R_1 is the average distance. Less typical samples will have a larger d_r. As a larger m and a smaller d_r indicates a confident prediction, we found it is convenient to define a confidence score as $C = m/d_r$. To make confident prediction, we require that the score is larger than a certain threshold (0.08-0.15).

In the example shown in Fig.1B, the new sample is confidently classified as type 1 if C is greater than 0.08. If C fails to reach the threshold value, a 'null' prediction is made. This may be caused by a small m, which indicates that the new sample is almost equally similar to the two best matches. Or, this may also happen if d_r is much larger than 1. In this case, even though the new sample are more similar to the prototype G_1, it deviates from most samples of this kind in the training set so significantly that the difference can no longer be explained in terms of random variance. In addition, the raw Pearson's correlation coefficient between the new sample and prototype G_1 should be larger than 0.2. All together, these criteria help the algorithm avoid false positives. The simplicity of the PM algorithm makes it convenient to impose various common-sense based constraints for making predictions without the significant influence on false negative error rates.

These measures of prediction confidence are chosen empirically. More rigorously, one could assume that the distances to a prototype follow a Gaussian distribution. Thus the mean and standard deviation could be used to evaluate the likeliness that a new sample with distance d_i belongs to a group. Such P values can be calculated for each of the prototypes and the new sample is assigned to the one with the smallest P value. Again, both the absolute P value and the margin should be taken into account to avoid false positives. However, since the number of biological replicates within each cancer type is usually very limited, the mean and standard deviation may not be very reliable. Such approach is not used in the present study. Rather, we used the empirical formulae that are believed to be more robust for small sample size.

For comparison, we also used the closely related KNN and SVM. KNN has many variations in the way that prototypes are chosen from training data and in the ways that votes are weighted[13]. Here a new sample is compared with all the samples in the training dataset. (Unlike PM, KNN uses all the samples in the training dataset as prototypes.) Then the 8-10 nearest neighbors vote with a weight of $1/r$, where r is the rank. The class that receives most votes wins. Confidence is simply characterized by the margin in the percentage of vote. The threshold to make diagnosis prediction is set to as high as 80%, which requires that most of the k neighbors should belong to one class. For SVM, we used an implementation of SVM-FU (www.ai.mit.edu/projects/cbcl) developed by Ryan Rifkin.

2.3 Validation: sensitivity vs. reliability

There are three kinds of predictions errors. A false negative error refers to the case that a null prediction is made for samples that actually belong to the tumor types in the training set. On the contrary, a false positive error corresponds to the case that a positive prediction is made for those samples that do not belong to any of the tumor types. Also, samples that belong to one of the tumor types may be misclassified. Such cases are usually rare. The performance of classification system should be evaluated with regard to both false negative and false positive.

The false negative error rate is usually evaluated through two tests. In leave-one-sample-out cross validation (LOSOCV), each sample in the training set is withheld and used to test the performance of the classifier trained on the remaining samples. In the 'positive test', independent samples that belong to the training subtypes are presented to a classifier. These samples should be confidently assigned into one of the classes.

To evaluate the false positive error rate, we introduce a 'null test'. In this test, we present a classifier samples that do not belong to any of the categories in the training dataset. Such samples can be, for example, normal tissues or those from other organs. For these samples, a reliable algorithm should produce a 'null' prediction because they should not be assigned to any of the subtypes known to the classifier. Otherwise, a false positive error is registered.

Sometimes, however, null test is impossible due to the lack of samples. An alternative procedure called leave-one-class-out cross validation (LOCOCV) is used. Withholding all the samples that belong to one tumor subtype, we train a classifier with the remaining samples. Then the classifier is tested against false positive error by presenting the samples that are left out. Basically it is a generalization of LOSOCV. The difference is that one withholds a cancer subtype instead of a sample. Note that LOCOCV is only applicable to larger datasets with more subtypes so that the elimination of one cancer subtype does not influence significantly the performance of the classifier. In the next section we apply this procedure to a dataset of 11 cancer types.

2.4 Outliers in the training dataset

The training dataset may contain a small number of outliers due to a variety of reasons such as sample preparation, array experiment, clinical diagnosis, etc. A small number of outliers in the training dataset could seriously degrade the performance of classifiers. As indicated in Fig. 1A, we eliminate such samples from the training dataset according to LOSOCV. We reasoned that the training dataset should be consistent with itself. In our calculation, a sample is considered an outlier if (a) it is misclassified with a high C value when it is not used for training and (b) this single sample exerts un-proportionally large influence on the overall classifier. But one should be very careful with the elimination of samples because the effect of eliminating different samples might be inter-dependent. Additionally, the total number of samples to be eliminated should be kept small (less than 5%). For the detection of outliers, it would be helpful to examine the dataset with some outside programs such as hierarchical clustering and data visualization algorithms.

3 Datasets and Results

Leukemia dataset and the hidden false-positives. The main purpose of the first case of application is to test the efficiency of our statistical feature selection scheme and the robustness of the PM against false positives. For the training dataset, we used the leukemia dataset [1]. This dataset contains expression patterns of samples for acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). As ALL samples can be further divided into two groups: T-cell lineage and B-cell lineage, we consider three subtypes in this dataset. There are 38 samples in the training dataset(11 AML, 19 B-ALL and 8 T-ALL), and 34 independent samples for positive-test. Microarrays used in the experiment are Affymetrix HuFL which contains probes for 6817 human genes. Expression level of a gene is characterized by the average difference score of multiple match and mismatch probe pairs.

To test the false positive rates, we incorporated datasets from several laboratories as the HuFL chip has been widely used and many datasets are available to the public. Our null-test data includes an ovary dataset [6], a dataset of stomach and liver tissue samples, and a variety of other samples from

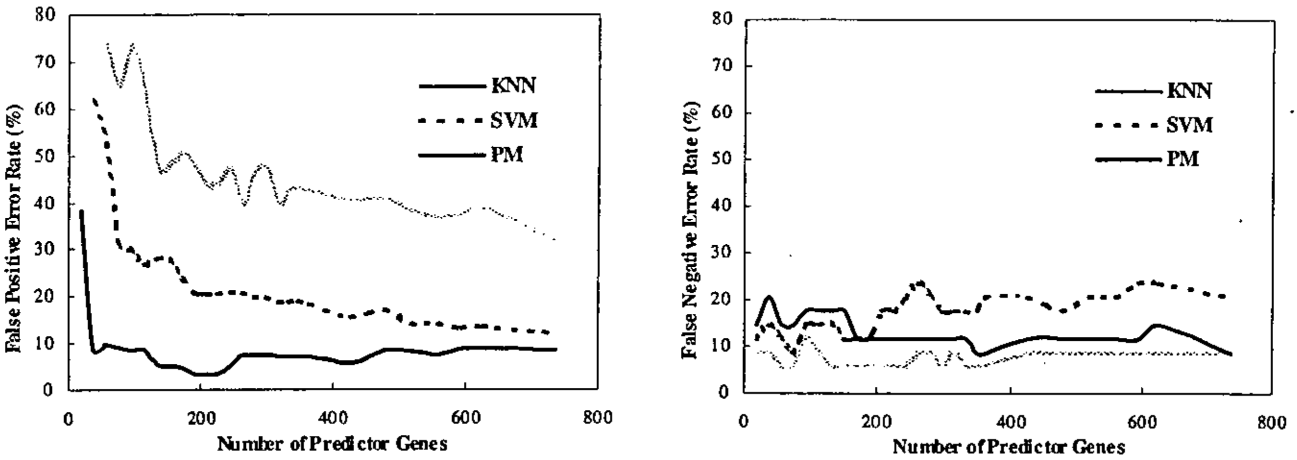


Figure 2: The change of false positive (left) and false negative (right) error rates with the number of genes used for prediction. KNN has the highest false positive error rates and the lowest false negative error rates. For SVM, although we raise the prediction threshold so high that its false negative error rate increases to about 20%, its false positive error rates are still higher than that of PM. This suggests that SVM is intrinsically vulnerable to false positives. The performance of PM are reasonable in both kinds of errors.

our collaborators. These 239 samples are of various origins; they can be any types of human tissues except AML and ALL. See supplementary information for more information about all those samples. Expression scores on each array are normalized to have the same mean and standard deviation to make the data from different laboratories compatible.

In pre-processing, we first eliminate those genes that do not change significantly by requiring that the difference and ratio between the maximum and minimum expression level be larger than 300 and 2, respectively[1]. We also require that the standard deviation and its ratio to mean value be larger than 100 and 0.1. Those genes with a Kruskal-Wallis H score smaller than 9.21 are eliminated as their expression profiles do not correlate with tumor distinction with a statistical significance level P=0.01. The expression levels of the remaining 736 genes are log-transformed. Normalization is done simply by dividing the raw data by the length of a gene's expression vector so that each gene is characterized by a unitary vector. We found that the classification accuracy can be significantly degraded if we follow the popular way of normalization that makes all genes have the same variance.

According to their similarities, these genes are divided into 20 groups by K-means clustering. From each group we selected a small number of genes and construct a feature set. Based on these selected genes, prototype matching is used to make predictions on new samples. The threshold for the prediction confidence C is set to 0.15. By changing the number of genes selected from each cluster, the changes of false negative and false positive error rates are plotted in Fig. 2.

Surprisingly, a large difference is observed in the false positive rates of different classification methods. When less than 100 genes are used, KNN could have a false positive error rate as high as 50%. SVM also has a relatively high error rate of about 20%. On the contrary, PM has an error rate smaller than 10%.

Even with as few as 19 predictor genes, most samples in the positive-test can be correctly classified. This could misleadingly suggest the use of small feature set. But null-test indicates that the false positive rates could be as high as 92% for KNN, 89% for SVM and 38% for PM. Therefore, whatever the classification algorithms, it is important to include several hundreds of genes in the feature set. This shows the importance of null-test: those seemingly irrelevant datasets serve as a background based on which we can tell whether a feature set enables the unique definition of expression prototype for a certain cancer type.

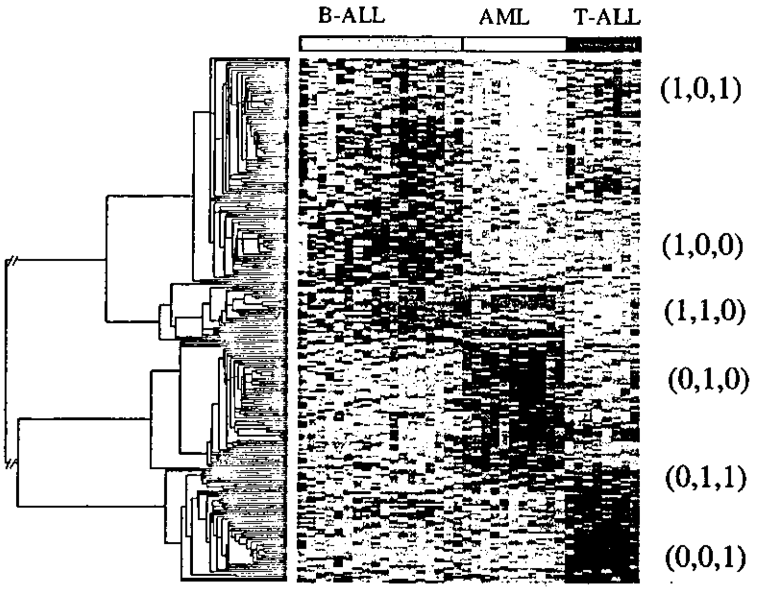


Figure 3: A set of 227 genes chosen for classification of B-cell acute lymphoblastic leukemia (B-ALL), T-cell acute lymphoblastic leukemia (T-ALL), and acute myeloid leukemia (AML)[1]. Instead of searching the whole gene list for predefined expression patterns, we try to find all existing patterns that could be helpful in cancer classification. Black indicates high expression.

From Fig. 2, we also observed that the error rate of PM is not sensitive to the number of genes used in prediction, as long as the number is not too small. Optimizing the balance between two kinds of errors, we finally select 227 genes in our final feature set. In LOSOCV, most of the samples in the training set are correctly classified except 3 false negatives. In the positive and null test, PM has 4 false negatives (11.8%) and 8 false positives(3.3%).

Figure 3 gives these informative genes. The feature set contains all 6 possible alternative expression patterns in a 3-class problem. From top to bottom in the figure, there are genes of type $\{1,0,1\}$, $\{1,0,0\}$, $\{1,1,0\}$, $\{0,1,0\}$, $\{0,1,1\}$, $\{0,0,1\}$, with 1 representing high expression and 0 low expression. Unsurprisingly, at the top region of the figure we find a large number of genes that are shared by B-ALL and T-ALL. We include these genes because we believe they can help achieve a high signal-to-noise ratio. Such genes are ignored in the one-vs-all gene selection method used by [7, 5, 8], as only genes of type $\{1,0,0\}$, $\{0,1,0\}$, $\{0,0,1\}$ are selected.

This is further justified by Fig. 1 in supplementary information, in which the false positive error rates using two feature sets are given. With PM, the cluster-and-select method yields more reliable predictions, especially when smaller feature sets are used. This might be attributed to the fact that the new feature selection method includes some very informative genes that are ignored by the conventional method. When more than 500 genes are included in the feature set, error rates tend to be very close. When SVM is applied to these two feature sets, a similar tendency is observed. However, the difference in false positive error rate between feature selection procedures is subtle in comparison with that due to classification algorithms, especially when more genes are used for prediction.

Finally, Fig. 4 shows the distribution of all samples with regard to prediction parameters m and d_r. The x axis is the relative distance to the nearest prototype, while the y axis represents the margin of this prototype over the second nearest one. These two parameters have intuitively simple meanings that could be more easily understood by biologists than parameters like vote percentage or the output of artificial neurons. Because both x and y axes relate to the distances of samples, we refer to such a plot as distance-distance (DD) plot. Most samples in the training and the positive test dataset are located in regions with a large m and small d_r. On the contrary, samples in the null-test dataset have a small m and a large d_r. The decision line for confident prediction $m = td_r$ is also drawn. Here

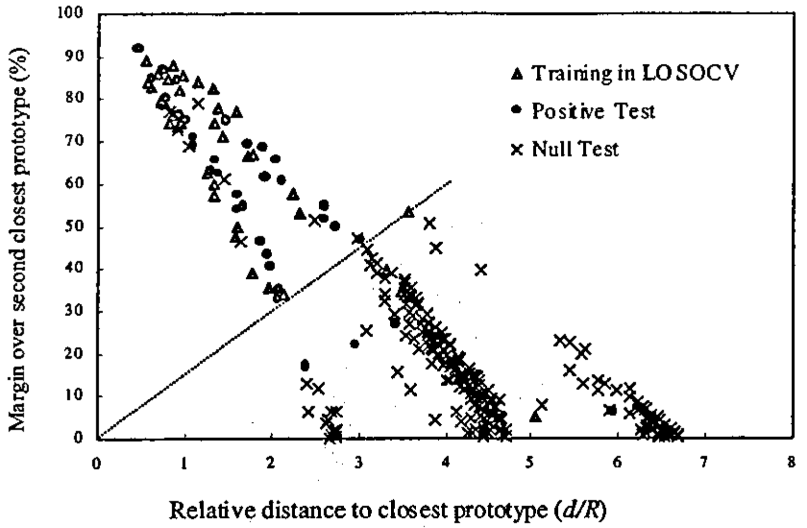


Figure 4: Distribution of samples plotted by two prediction parameters d_r and m. Parameter d_r is the relative distance to the nearest prototype and m is the margin over the second nearest. Positive predictions are made for the samples that lie above the dashed line.

$t = 0.15$ is the slop. In general, the choice of the t can be made according to a plot like Fig. 4, which helps to balance false negative and false positive error.

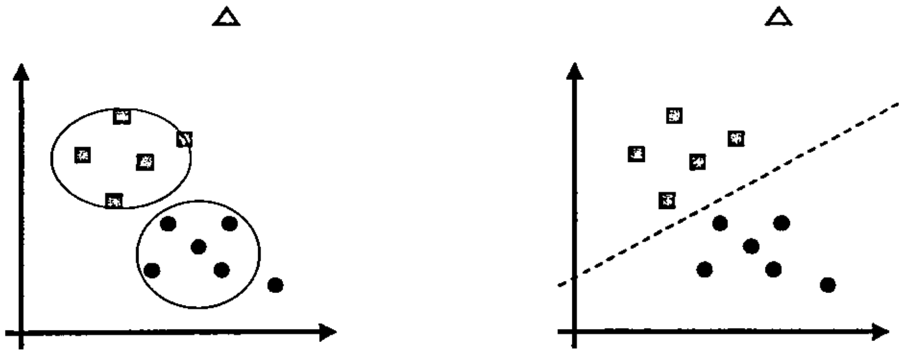
Extensive testing on other datasets. We then tested the PM algorithm on several other datasets including a lymphoma dataset[4], a SRBCT dataset [5], and the dataset of Su et al.[7]. Prediction results are summarized in table 1. More information is available in the Supplementary Information. The classification scheme proposed in this paper has a relatively lower rate of both false positive and false negative error in these datasets.

Table 1: Summery of four datasets and the performance of PM algorithm. Given in parentheses are the number of misclassified cases observed in leave-one-sample-out-cross-validation (LOSOCV) or independent test. Note that the false negative error rate is calculated according to both LOSOCV and positive test. For the dataset of Su *et al.*, which no independent data for null test is available, a leave-one-class-out-cross-validation (LOCOCV) procedure is employed.

	# Tumor	Samples size			False	False	# Genes
Dataset	types	Traning	Positive test	Null test	negative	positive	used
Leukemia [1]	3	37 (3)	34 (4)	239 (8)	11.8%	3.3%	227
Lymphoma [4]	3	40 (7)	26 (6)	27 (2)	19.7%	7.4%	328
SRBCT [5]	4	63 (6)	20 (2)	6 (0)	9.6%	0%	390
Su *et al.* [7]	11	97 (13)	74 (14)	—(12)	15.8%	12.4%*	400

4 Discussion

There are a wealth of statistical and machine learning tools that could be useful for the classification of cancers. How to pick up the right tools and integrate them is an important issue. Such choice should be made based on knowledge of underlying computational principles. Classification algorithms like SVM and KNN define a hyperplane or hypersurface according to which a multidimensional space of samples are divided into two or more regions (Fig.5). Implicitly, it is assumed that all samples presented to the classifier belong to at least one of the predefined tumor types. This might be true



A. Prototype matching (PM)

B. Support vector machine (SVM)

Figure 5: Prototype matching (PM) and support vector machine(SVM) belong to two different paradigms for pattern recognition. While SVM makes prediction for all samples that are far enough from the hyper-plane, PM makes positive predictions only if a new sample is sufficiently similar to a prototype. K-nearest neighbor method is very similar to SVM. As a result, classifiers like PM tend to have more false negatives (e.g. the sample in the bottom right) while those like SVM and KNN may suffer seriously from a high false positive rates(e.g. the sample marked by a triangle at the top).

in some classification tasks such as metastatic vs. non-metastatic tumor[9], or curable vs. incurable DLBCL patients[10]. These are 'true' binary problems, in which SVM and KNN can make accurate predictions. But 'pseudo' binary classification tasks are more freqent: there could exit a third class missing in both training and test dataset. Clinically, the existence of new subtypes of cancers are always possible and it is very difficult to obtain a 'complete' training dataset as required by SVM and KNN. SVM and KNN may have high false positive rates when presented with samples of novel tumor types. This is confirmed in this study (Fig.2).

Unlike SVM and KNN, PM defines a closely bounded region in the multidimensional space to represent each tumor subtypes(Fig.5). There is a large rejection zone that a 'null' prediction will be outputted. The uniqueness of each subtypes is recognized by an expression prototype. Although PM can have a slightly higher false negative rate, false positive error is found to be much lower. Avoiding false positives is essential in the process of discovering new cancer subtypes. Therefore, we believe PM and methods alike might be more suitable for cancer classification.

Our results also give some hints to the question of how many predictive genes should be used in cancer classification. With the inclusion of more genes, we found that false positive errors decrease accordingly. But the opposite tendency is often observed for false negative error. Therefore, the optimal choice should be made by seeking a balance. This could be done by the minimization of the total error rate.

The searching of differentially expressed genes in two or more groups is one of the fundamentally important tasks in the bioinformatics of gene expression analysis. Besides cancer classification, the cluster-and-select feature selection procedure proposed here might be useful in this context. The procedure is able to detect different patterns of gene expression in multiple groups.

To select features from the highly redundant measurements in expression profiling, we employed k-means clustering in addition to a statistics score. Unsupervised classification techniques themselves are used in the process of feature selection for the purpose of supervised classification. This strategy might be useful in other pattern recognition tasks where redundant measurements are involved.

Acknowledgements. We thank Shogo Yamamoto, M. Kano, Jerome Piat, and Naohiro Shichijo for stimulating discussions. We are also greatly indebted to Todd R. Golub for helpful comments on our manuscript and Jiang Fu for proof-reading. Critical comments from the reviews also helped greatly to improve the presentation of this paper.

Supplementary Information. Supplementary information is avalible at web site of Genome Informatics, also available at: www.race.u-tokyo.ac.jp/ xge/cancer/.

References

[1] Golub, T. R. , *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537, 1999.

[2] Bhattacharjee A., *et al.*, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.*Proc. Natl. Acad. Sci. USA* 98:13790-13795, 2001.

[3] Ramaswamy,S., & Golub,T. R. DNA microarrays in clinical oncology. *J. CLIN. ONCOL.* 20:1932-1941, 2002.

[4] Alizadeh, A.A.*et al.* The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes.*Nature (London)* 403, 503-511, 2000.

[5] Khan, J., *et al.*, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7:673-679, 2001.

[6] Welsh, J. B., *et al.*, Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl. Acad. Sci. USA* 98:1176-1181, 2001.

[7] 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. & Hampton, G. M. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research* 61:7388-7393, 2001.

[8] Ramaswamy S., *et al.*, Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* 98:15149-15154, 2001.

[9] Van 'T Veer L. J., *et al.*, Gene expression profiling predicts clinical outcome of breast cancer. *Nature (London)* 415:530-536, 2002.

[10] Shipp, *et al.*, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8:68-74, 2002.

[11] Dudoit S., Fridlyand J. & Speed T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97:77-87, 2002.

[12] Heydorn, R. P., IEEE Trans. C-2:1051-1054, 1971.

[13] Duda, R. O., Hart, P. E., & Stork, D. G., *Pattern classification* (Wiley, New York, NY), 2001.

[14] Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863-14868, 1998.

[15] Yeang, *et al.*Molecular classification of multiple tumor types. *Bioinformatics* 17 Suppl., S316-S322, 2001.

Supplementary Information

Reducing false positives in molecular pattern recognition

Xijin Ge[†], Shuichi Tsutsumi[†], Hiroyuki Aburatani[†], and Shuichi Iwata[*]
[†] *Genome Science Division, Research Center for Advanced Science and Technology(RCAST),*
[*] *Department of Quantum Engineering and Systems Science, School of Engineering,*
The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan

More information is available at http://www.race.u-tokyo.ac.jp/~xge/cancer/

1 The parameters (M and S)in feature selection

K-means clustering is employed to increase the signal-to-noise ration in feature selection. One of the difficult problems in clustering analysis is to decide how many clusters is appropriate for a certain dataset. Although there are some complicated methods that can make approximate recommendations, in most of the cases this problem is solved by trial-and-error. Similar to the detection of outliners, representation produced by hierarchical clustering and multidimensional scaling (MDS) can be used as a as a guidance.

In our calculation, we divide all the genes that passed the H test into M clusters and select S genes from each of them. We studied the effects of different M and S on classification error rates in the leukemia dataset. The result of false positive and false negative error are showing in Table 1 and 2, respectively. As indicated in the tables, false positive error rates can change from 3.35% to 10.04% under different parameters, all lower than what we could expect using SVM. False negatives can change from 8.82% to 17.65%. Although error rates do vary in different conditions, but our classification schemes are relatively stable.

2 Lymphoma dataset

We choose the dataset of Ref. [4] because it contains not only samples of three most prevalent adult lymphoid malignancies but also samples of purified normal lymphocyte subpopulations that can be used in

Table 1: False positive error rates with different choices of parameters observed in the leukemia dataset.

	Number of clusters(M)		
genes (S)	15	20	30
5	10.04%	8.37%	9.62%
8	10.46%	5.86%	6.69%
10	8.37%	3.35%	6.69%
12	7.11%	3.35%	7.11%
15	7.11%	7.53%	7.53%
20	9.21%	7.11%	8.37%
50	7.95%	7.95%	7.95%

Table 2: False negative error rates with different choices of parameters observed in the leukemia dataset.

	Number of clusters(M)		
genes (S)	15	20	30
5	14.71%	17.65%	17.65%
8	17.65%	17.65%	11.76%
10	14.71%	14.71%	14.71%
12	17.65%	14.71%	11.76%
15	14.71%	14.71%	14.71%
20	11.76%	8.82%	11.76%
50	11.76%	17.65%	14.71%

null test. The three types of malignancies are diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), and chronic lymphocytic leukemia (CLL). The training dataset contains 23 DLBCL samples, 8 FLs and 9 CLLs. We have 26 samples for the positive-test and 27 samples for null test. The dataset is obtained by using the 'Lymphochip': a specialized DNA microarray that contains probes for the genes that are preferentially expressed in lymphoid cells and for the genes with known or suspected roles in processes important in immunology or cancer.

Beginning with the 4026 array elements that passed the filter in the original study[4], we eliminated 3239 genes through a preprocessing procedure (similar to the one used in the leukemia dataset) and a Kruskal-Wallis H test with a minimum P value of 0.025 (H=7.38). The final feature set contains 328 genes (Fig. 3). The threshold for prediction confidence is 0.15. All except two samples in the null test are classified as 'null'. Six false negatives are observed in the 26 samples in the positive-test.

3 SRBCT dataset

The third dataset is from Khan *et al.* [5]. It contains four subtypes of SRBCT, namely neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). The training set contains 12 samples for NB, 20 for RMS, 23 for EWS, and 8 for Burkitt lymphomas (BL, a subset of NHL). There are 25 samples in the independent dataset, among which 6 are normal skeletal muscle and other kinds of tissues that can be used in the null test. Microarray experiment is based on NHGRI protocol. Different from the leukemia data, relative red intensity, which is the ratio of the mean intensity of a spot to overall average intensity, is used as a measure for expression level.

Of the 2308 genes in the raw dataset, 841 passed the H test with P<0.05 (H>7.81). Using similar preprocessing and cluster-and-select method, we choose a final feature set of 390 genes ($M = 20$, $S = 6$). It is found this set of genes includes 75 of the 96 genes selected by Khan *et al.*[5]. Among those common genes is the fibroblast growth factor receptor 4 (FGFR4) that is activated in RMS samples. This gene has an H score as high as 39.93. As indicated by Fig. 4, hierarchical clustering using these genes are consistent with clinical diagnosis. The threshold for making confident prediction is set as $C = 0.08$. All of the 6 control samples are correctly predicted as 'null'. There is one false-negative prediction in the independent positive test.

4 Dataset of Su *et al.*

To demonstrate the efficiency of our algorithm on larger datasets, we consider the expression data of Su *et al.* [7]. This dataset contains 100 samples representing 11 tumor types, namely, prostate, breast, lung, ovary, colorectum, kidney, liver, pancreas, bladder/ureter and gastroesophagus. In addition, there are 74 independent samples for positive test. The expression levels of a total of 12533 probes are obtained by using Affymetrix Hu6800 and Hu35KsubA microarray.

We eliminated 3 outlier samples in the training set. As many as 5935 genes passed the H test with P$\leq$0.05 (H$\geq$18.31). These genes are clustered into 50 groups. Selecting 8 genes from each group, we obtained a feature set of 400 genes(Fig. 5). Thirteen false negatives are observed in LOSOCV with a threshold confidence level of $C \geq 0.15$. Null predictions are produced for fourteen of the 74 independent samples. Altogether, the false negative error rate is 15.8%, which is slightly higher than 13.5% reported in the original study with SVM[7].

The LOCOCV procedure described in section 2.3 is employed to evaluate the false positive error rate. We withhold one of the 11 tumor types and train classifiers to distinguish the rest 10 types. Then the unused samples are presented to the classifier to see if it misclassifies these samples as any of the trained tumor types. The procedure is repeated for each of the 11 tumor types, accumulating false positive cases. Totally, 12 such errors are observed, leading to a false positive error rate of 12.4%. With the same set of predictor genes, SVM misclassified 14 samples (14.4%). With the one-vs-all feature set, PM and SVM have a false positive error of 16.8% and 18.7%, respectively. Therefore, the LOCOCV makes it possible to test classifiers against false positive errors even without independent control samples.

5 List of genes used for classification

The list of genes that are selected by our algorithm are given in our web page (http://www.race.u-tokyo.ac.jp/~xge/cancer/). Information are availabe for following datasets:

- Leukemia dataset
- Lymphoma dataset
- SRBCT dataset
- Dataset of Su et al.

References

[1] Golub, T. R. , Slonim, D. K. , Tamayo, P. , Huard, C. , Gaasenbeek, M. , Mesirov, J. P. , Coller, H. , Loh, M. L. , Downing, J. R. , Caligiuri, M. A. , *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

[2] Bhattacharjee A., Richards W.G., Staunton J., Li C., Monti S., Vasa P., Ladd C., Beheshti J., Bueno R., Gillette M., Loda M., Weber.G., Mark E.J., Lander E.S., Wong W., Johnson B.E., Golub T. R., Sugarbaker DJ, Meyerson M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.*Proc. Natl. Acad. Sci. USA* **98**, 13790-13795.

[3] Ramaswamy,S., & Golub,T. R. (2002) DNA microarrays in clinical oncology. *J. CLIN. ONCOL.* **20**, 1932-1941.

[4] 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., *et al.* (2000) The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes.*Nature (London)* **403**, 503-511.

[5] Khan, J., Wei., J. S., Ringér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7**, 673-679.

[6] Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., Lockhart, D. J., Burger, R. A. & Hampton, G. M (2001) Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl. Acad. Sci. USA* **98**, 1176-1181.

[7] 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. & Hampton, G. M. (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research* **61**, 7388-7393.

[8] Ramaswamy S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M. Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P. Poggio, T., Gerald, W., Loda, M., Lander, E. S., & Golub, T. R., (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* **98**, 15149-15154.

[9] Van 'T Veer L. J., Dai H., Van de Vijver M. J., He Y. D., Hart A. A. M., Mao M., Peterse H. L., Van der Kooy K., Marton M. J., Witteveen A. T., Schreiber G. J., Kerkhoven R. M., Roberts C., Linsley P. S., Bernards R., and Friend S. H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature (London)* **415**, 530-536.

[10] Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C. & Golub, T. R. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* **8** 68-74.

[11] Dudoit S., Fridlyand J. & Speed T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**, 77-87.

[12] Heydorn, R. P. (1971). IEEE Trans. **C-2**, 1051-1054.

[13] Duda, R. O., Hart, P. E., & Stork, D. G., (2001) *Pattern classification* (Wiley, New York, NY).

[14] Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868.

[15] Yeang, C.-H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., Reich, M., Lander, E. , Mesirov, J. & Golub, T. (2001) Molecular classification of multiple tumor types. *Bioinformatics* **17** Suppl., S316-S322.