

ARTICLE

Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis

Xue Wu Zhang^{*,1}, Yee Leng Yap², Dong Wei¹, Feng Chen^{1,3} and Antoine Danchin⁴

¹College of Light Industry and Food Sciences, South China University of Technology, Guangzhou, China; ²Department of Microbiology, The University of Hong Kong, Pokfulam, Hong Kong, China; ³Department of Botany, The University of Hong Kong, Pokfulam, Hong Kong, China; ⁴Genetics of Bacterial Genomes, Institut Pasteur, CNRS URA2171, Paris, France

The precise diagnosis of cancer type based on microarray data is of particular importance and is also a challenging task. We have devised a novel pattern recognition procedure based on independent component analysis (ICA). Different from the conventional cancer classification methods, which are limited in their clinical applicability of cancer diagnosis, our method extracts explicitly, by ICA algorithm, a set of specific diagnostic patterns of normal and tumor tissues corresponding to a set of biomarkers for clinical use. We validated our procedure with the colon and prostate cancer data sets and achieved good diagnosis (>90%) on the data sets studied here. This technique is also suitable for the identification of diagnostic expression patterns for other human cancers and demonstrates the feasibility of simple and accurate molecular cancer diagnostics for clinical implementation.

European Journal of Human Genetics advance online publication, 5 October 2005; doi:10.1038/sj.ejhg.5201495

Keywords: microarray; cancer; ICA; diagnosis; pattern; biomarkers identification

Introduction

Conventional diagnosis of cancer relies on macro- and microscopic histology and tumor morphology. This methodology is somewhat subjective and depends on highly trained pathologists. Furthermore, there is a wide spectrum in cancer morphology and many tumors are atypical or lack morphologic features that are useful for different diagnosis.¹ Recent years witnessed an increasing interest in changing the basis of tumor classification from morphologic classification to molecular genetics-based classification. The rapid development of microarray technologies that can simultaneously assess the expression level of

thousands of genes offers the promise of precise, objective and systematic human cancer classification using molecular diagnosis. Many techniques have been used to analyze gene expression data and have demonstrated the potential power of expression profiling for tumor classification (see for review Simon *et al*²).

Independent component analysis (ICA) is a dimension reduction technique that uses the existence of independent factors in multivariate data and decomposes an input data set into statistically independent components. ICA can reduce the effects of noise or artifacts of the signal and is ideal for separating mixed signals.³ ICA has been used successfully in electroencephalographic (EEG), magnetoencephalographic (MEG) and functional magnetic resonance imaging (fMRI) data.^{4–6} Recently, Liebermeister⁷ used ICA for microarray analysis to extract expression modes of genes. Lee and Batzoglou⁸ conducted a systematic analysis of the applicability of ICA to microarray data. Moreover, a

*Correspondence: Dr XW Zhang, College of Light Industry and Food Sciences, South China University of Technology, Wushan Road, Tianhe Area, Guangzhou 510641, China.

Tel: +86 20 87113849; Fax: +86 20 87113843;

E-mail: snow_dance@sina.com

Received 18 June 2005; revised 4 August 2005; accepted 11 August 2005

recent report⁹ indicated that ICA could improve the biological validity of the genes identified as differentially expressed in endometrial carcinoma, compared to other techniques such as Cyber-T (a Bayesian framework for the analysis of microarray expression data using *t*-test, <http://visitor.ics.uci.edu/genex/cybert/index.shtml>) and significance analysis of microarrays (SAM, <http://www-stat.stanford.edu/~tibs/SAM/>). In this study, we developed an ICA-based algorithm for classifying tissues on the basis of gene expression data. Different from previous methods, our method identified not only a set of biomarkers but also a set of specific diagnosis patterns of normal and tumor samples corresponding to these biomarkers. Using this method, we analyzed colon and prostate cancer data and demonstrated that this method outperformed previous studies.

Data and methods

Microarray data sets

The gene expression data sets from colon and prostate cancers were investigated in this study. For colon cancer, the data set is the expression profiles of 2000 genes using Affymetrix Hum6000 arrays in 22 normal and 40 colon cancer tissue samples¹⁰ (the normalized data set can be downloaded at <http://microarray.princeton.edu/oncology/affydata/index.html>). For prostate cancer, three data sets were used, which are the expression profiles of 12 600 genes using Affymetrix U95Av2 arrays. The first data set consists of 50 normal and 52 prostate cancer tissue samples; the second data set includes 10 nonrecurrent and eight recurrent prostate cancer tissue samples (Department of Adult Oncology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA).¹¹ The third data set is an independent data set of nine normal and 25 tumor prostate samples from another laboratory (Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA).¹² The normalized data sets can be downloaded at <http://www-genome.wi.mit.edu/MPR/Prostate>.

Mathematical framework of ICA

Given a microarray data set $X = (x_{ij})_{m \times n} = (x_1, \dots, x_m)^T$ (T means transpose) with m rows of genes and n columns of samples (ie n different experimental conditions), each element x_{ij} in the matrix X corresponds to the i th gene's expression level in the j th sample. If the expressions of m genes are governed by k independent biological processes, such as ribosome biogenesis, cell cycle, etc, then $S = (s_1, \dots, s_k)^T$ ($k \leq m$). We assume that the expression of each gene x_i ($i = 1, \dots, m$) is a linear combination of the k independent biological processes s_j ($j = 1, \dots, k$) with some unknown mixing coefficients a_{ij} : $x_i = \sum_j a_{ij}s_j$, written in the form of matrix representation

$$\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mk} \end{pmatrix} \begin{pmatrix} s_1 \\ \vdots \\ s_k \end{pmatrix}$$

or

$$X = AS$$

A is called the mixing matrix and S is called source signals. The goal of ICA is to find a matrix W that satisfies the transformation equation

$$Y = WX = WAS$$

W is called separating matrix and $Y = (y_1, \dots, y_k)$, called independent components, has statistically independent components. Generally speaking, Y is a close approximation of source signal S ; if $W = A^{-1}$, it achieves perfect reconstruction $Y = S$. To find such a matrix W , an important assumption is that at most one source signal has a Gaussian distribution. This is not a problem for analyzing biological data based on the fact that the most typical Gaussian source is random noise and biological processes are expected to be highly nonrandom, that is, non-Gaussian; for example, in the regulation of gene expression, a set of relevant genes are sharply affected and most other genes are relatively unaffected.⁸

ICA-based diagnosis algorithm

Our statistical algorithm is a combination of the following sequential steps.

Data preprocessing Prior to further analysis, log₂ transformation was performed on the colon data. Because too many genes (12 600) are included in the prostate data, we first removed those genes whose expression level is less than 2, retaining the most significant 1662 genes, and then log₂ transformation was performed.

Sampling For diagnostic purposes, 50% of the samples are randomly selected from normal and tumor samples as the training data set and the remaining data constitute the test data set.

Extraction of independent components The FastICA program (<http://www.cis.hut.fi/projects/ica/fastica/>) was used here. If X_{normal} represents the normal training data set, we performed ICA on the transpose of data matrix X_{normal} and extracted one independent component IC_{normal} with dimension $m \times 1$ (corresponding to the largest one and accounting for 70% of the variance). Similarly, if X_{tumor} represents the tumor training data set, we performed ICA on the transpose of data matrix X_{tumor} and extracted one independent component IC_{tumor} with dimension $m \times 1$ (corresponding to the largest one and

accounting for 70% of variance). Here, after the experimental noise was reduced by ICA, we expected IC_{normal} and IC_{tumor} to represent the characteristic expression profile of all genes in normal samples and tumor samples, respectively, for extraction of two diagnostic patterns. The MATLAB code using the FastICA program was as follows:

$$IC_{\text{normal}} = \text{FASTICA}(X'_{\text{normal}}, 'lastEig', 1, 'numOfIC', 1)$$

$$IC_{\text{tumor}} = \text{FASTICA}(X'_{\text{tumor}}, 'lastEig', 1, 'numOfIC', 1)$$

Biomarkers selection/test data set validation We first calculate the ratios using independent component values (loads):

$$R^i = \frac{IC_{\text{tumor}}^i}{IC_{\text{normal}}^i}, i = 1, \dots, m$$

Then, we performed biomarker selection. Specifically, the selection procedure of a subset of biomarkers starts from a pair of genes with the smallest and the largest R values, for example, genes i and j , and the corresponding loads in two independent components IC_{normal} and IC_{tumor} constitute two discriminant vectors:

$$V_{\text{normal}} = \begin{pmatrix} IC_{\text{normal}}^i \\ IC_{\text{normal}}^j \end{pmatrix}, V_{\text{tumor}} = \begin{pmatrix} IC_{\text{tumor}}^i \\ IC_{\text{tumor}}^j \end{pmatrix}$$

Subsequently, the two vectors are used for discriminant analysis: taking a sample from the test data set (ie the remaining 50% of the data), the expression intensities of the above two genes constitute a test vector:

$$V_{\text{test}} = \begin{pmatrix} \text{intensity}^i \\ \text{intensity}^j \end{pmatrix}$$

Calculate its distance (this refers to the Euclidean distance between vectors) with the two discriminant vectors obtained above. If the distance between V_{test} and V_{normal} was smaller than the distance between V_{test} and V_{tumor} , then the sample is classified as normal sample, otherwise as tumor sample. After all samples from the test data set had been evaluated following this procedure, we checked whether the classification rate based on the two genes achieved any user-predefined classification accuracy (eg 90–100%). If so, both genes were selected as biomarkers, if not, we included one more pair of genes in the study, those having the second smallest and the second largest R values. We then checked whether the four genes could achieve the user-defined classification accuracy in the same way as above. If so, the four genes were selected as biomarkers, if not, the whole procedure was repeated till we reached a set of genes achieving the required classification performance (90–100%). They were then selected as biomarkers.

Leave-one-out crossvalidation In order to get an unbiased estimation of the error rate associated with the

method, a commonly used statistical approach,² leave-one-out (or jackknife) crossvalidation, was employed. This method involves randomly withholding one of the samples analyzed, including both training and test data sets, building a predictor based only on the remaining samples, and then predicting the class of the sample left out. The process is repeated for each sample, and the cumulative error rate is calculated. If the final cumulative error rate was <10%, the leave-one-out crossvalidation was considered completed, otherwise, we repeated the above biomarker selection process to get another subset of biomarkers and did crossvalidation again, till the error rate was less than 10%.

Diagnostic pattern After the above steps were completed, we obtained a final set of biomarkers that meet our two requirements: (a) the classification accuracy is >90% for the test data set; (b) the error rate is <10% in leave-one-out crossvalidation. The loads of these biomarkers in the original independent component obtained from normal (respectively, tumor) samples IC_{normal} (IC_{tumor}) constituted the diagnostic pattern for normal (respectively, tumor) tissues.

Diagnosis Given a sample from the test data set or other independent data set, we simply calculate its distance (Euclidean distance between vectors) to the two diagnostic patterns obtained above, then compare the two distances. If the sample under investigation is closer to the normal pattern than to the tumor pattern, then it is diagnosed as normal sample, otherwise as tumor sample.

The schematic illustration of ICA-based diagnosis procedure is shown in Figure 1.

Results

We first applied our algorithm to colon microarray data and obtained three diagnostic models (Table 1). The correct prediction rate ranged from 90 to 100%. This means that we could achieve 100% prediction accuracy using 10 genes. Among these, five are overexpressed in normal tissues: 1843 (Gelsolin), 1423 (Myosin regulatory light chain 2), 897 (Complement factor D), 1387 (Phosphoenolpyruvate carboxykinase) and 1635 (Vasoactive intestinal peptide (VIP)); the remaining five genes are upregulated in tumor tissue samples: 1810 (Pancreatic stone protein (PSP)), 1473 (Translational initiation factor 2), 1771 (Aspartyl-tRNA synthetase alpha-2), 1935 (Wee1 hu gene) and 1671 (Human monocyte-derived neutrophil-activating protein (MONAP)).

Secondly, we applied our algorithm to prostate cancer data. Similarly, we built three diagnostic models (Table 2). Interestingly, the fewer genes in model 3 achieve higher prediction accuracy as compared to model 2. This shows that we could use as few as four genes to completely

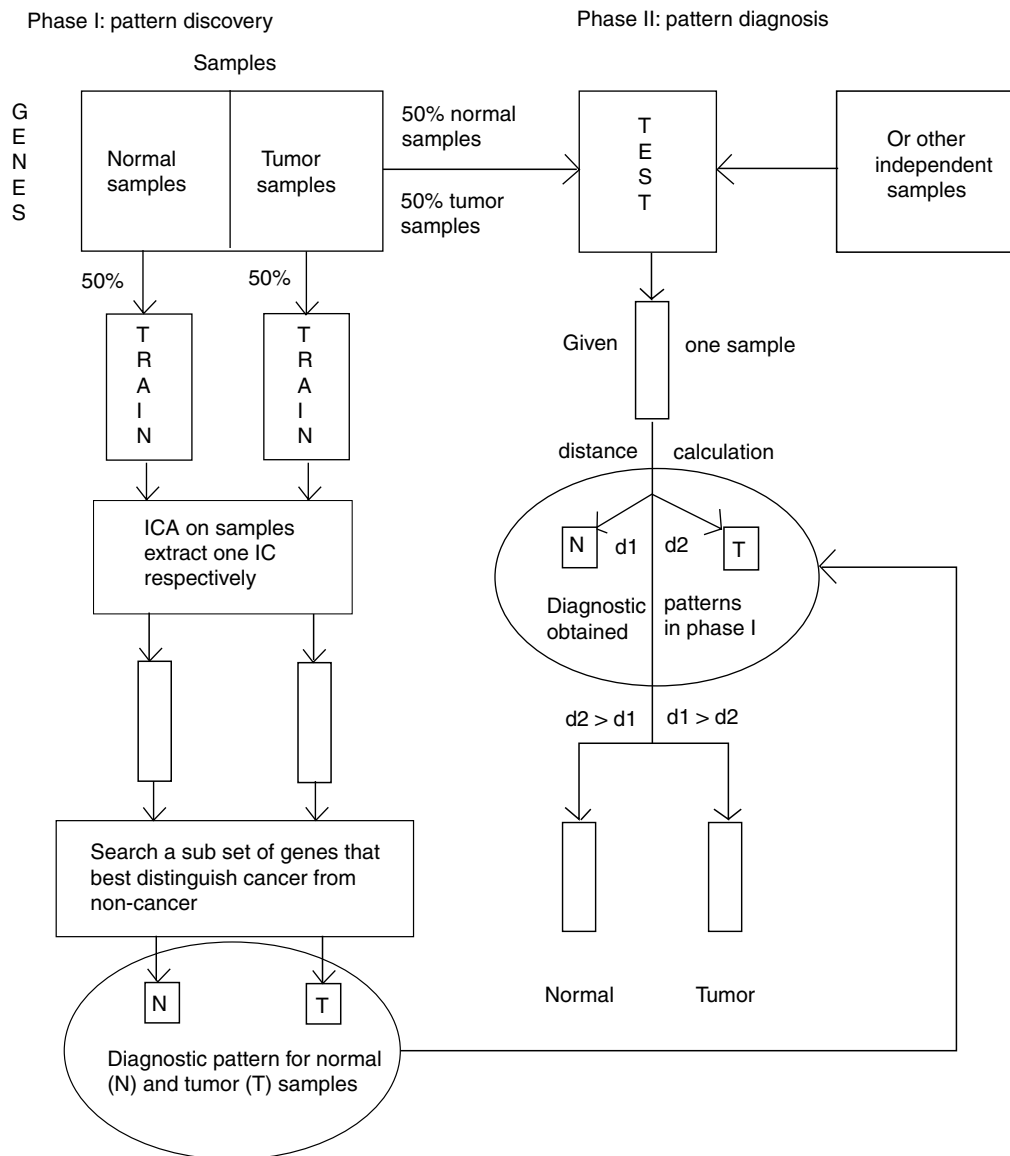


Figure 1 Scheme for ICA-based diagnostic method (see text for details).

diagnose test samples with 100% accuracy. These four genes included two downregulated genes in tumor samples (9850 (Human adipsin/complement factor D) and 6715 (an unknown gene from human melanocyte)) and two up-regulated genes in tumor samples (10875 (Human prostate carcinoma tumor antigen (PCTA-1)) and 6185 (Serine protease hepsin)).

To test the usability of our algorithm, we next applied the present three diagnostic models to prostate cancer to diagnose an independent data set of nine normal and 25 tumor prostate samples from Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA.¹² There was a nearly 10-fold difference in the overall

microarray intensity between this data set and the previous data set. There was a small difference in the number of genes in the study although they were using the same Affymetrix Hum95Av2 array:¹¹ 12 600 genes in the previous data set and 12 626 genes in the present data set; so the 26 extra genes were removed (see Supplementary Information). As an illustration, the diagnostic process using model 2 in prostate cancer (Table 2) was performed as follows:

(a) Given a sample from the independent data set of nine normal and 25 tumor prostate samples, for example, No. 4 normal sample, we included the expression values of only six genes in model 2 (9850, 11052, 4525, 5398, 4483, 6185) (Table 2) as a test vector:

Table 1 ICA-based diagnosis for colon cancer

<i>Model 1 (four genes)</i>		
<i>Gene</i>	<i>GenBank no.</i>	<i>Gene name</i>
1423	J02854	Myosin regulatory light chain 2, smooth muscle isoform (human); contains element TAR1 repetitive element
1635	M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds
1582	X63629	Homo sapiens mRNA for p cadherin
1671	M26383	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds

Diagnostic expression patterns for the above four genes

Normal	[127.6722,49.4295,16.6206,22.7769]
Tumor	[46.7113,23.7047,42.7040,71.5064]

Diagnostic result using the above patterns on the test data set

90.90% for normal, 90% for tumor, 90.30% for normal and tumor, 91.90% for crossvalidation

Model 2 (eight genes)

<i>Gene</i>	<i>GenBank no.</i>	<i>Gene name</i>
1843	H06524	Gelsolin precursor, plasma (human)
1423	J02854	Myosin regulatory light chain 2, smooth muscle isoform (human); contains element TAR1 repetitive element
1635	M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds
897	H43887	Complement factor d precursor (Homo sapiens)
1679	X53586	Human mRNA for integrin alpha 6
1771	J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds
1582	X63629	Homo sapiens mRNA for p cadherin
1671	M26383	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds

Diagnostic expression patterns for the above eight genes

Normal	[62.4920,104.6407,46.8053,135.4484,24.3357,23.5034,20.0135,25.1781]
Tumor	[30.2676,46.8702,25.0354,65.2678,47.2061,46.1609,38.3034,63.5447]

Diagnostic result using the above patterns on the test data set

100% for normal, 95% for tumor, 96.80% for normal and tumor, 91.90% for crossvalidation

Model 3 (10 genes)

<i>Gene</i>	<i>GenBank no.</i>	<i>Gene name</i>
1843	H06524	Gelsolin precursor, plasma (human)
1423	J02854	Myosin regulatory light chain 2, smooth muscle isoform (human); contains element TAR1 repetitive element
897	H43887	Complement factor d precursor (Homo sapiens)
1387	L05144	Phosphoenolpyruvate carboxykinase, cytosolic (human); contains Alu repetitive element; contains element PTR5 repetitive element
1635	M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds
1810	M27190	Homo sapiens secretory pancreatic stone protein (PSP-S) mRNA, complete cds
1473	R54097	Translational initiation factor 2 beta subunit (human)
1771	J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds
1935	X62048	Homo sapiens Wee1 hu gene
1671	M26383	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds

Diagnostic expression patterns for the above 10 genes

Normal	[72.3189,124.4136,146.2203,167.9398,52.1177,35.6766,28.7323,24.7062,14.4821,24.7422]
Tumor	[33.9585,55.5687,67.7242,76.8941,28.8660,80.8570,64.7676,55.0130,31.3784,79.4569]

Diagnostic result using the above patterns on the test data set

100% for normal, 100% for tumor, 100% for normal and tumor, 91.90% for crossvalidation

The ICA-based algorithm (see text for details) was sequentially performed three times on a colon cancer data set, and the expression profiles of 2000 genes using Affymetrix Hum6000 arrays in 22 normal and 40 colon cancer tissue samples¹⁰ and three diagnostic models (models 1–3) were obtained.

Table 2 ICA-based diagnosis for prostate cancer*Model 1 (four genes)*

Gene	GenBank no.	Gene name
9850	M84526	Human adipsin/complement factor D mRNA, complete cds
4525	Y16961	Homo sapiens mRNA for KET protein
9131	M14083	Human beta-migrating plasminogen activator inhibitor I mRNA, 3 end
6185	X07732	Human hepatoma mRNA for serine protease hepsin

Diagnostic expression patterns for the above four genes

Normal	[97.8025,21.2944,11.1016,22.9815]
Tumor	[19.3706,8.1072,17.1805,58.8879]

Diagnostic result using the above patterns on the test data set

92% for normal, 92.30% for tumor, 92.20% for normal and tumor, 90.20% for crossvalidation

Diagnostic result using the above patterns on an independent data set from another laboratory

100% for normal, 84.00% for tumor, 88.20% for normal and tumor

Model 2 (six genes)

Gene	GenBank no.	Gene name
9850	M84526	Human adipsin/complement factor D mRNA, complete cds
11052	J03242	Insulin-like growth factor 2
4525	Y16961	Homo sapiens mRNA for KET protein
5398	X70940	Homo sapiens mRNA for elongation factor 1 alpha-2
4483	AJ130733	Homo sapiens mRNA 2-methylacyl-CoA racemase
6185	X07732	Human hepatoma mRNA for serine protease hepsin

Diagnostic expression patterns for the above six genes

Normal	[84.3491,50.3947,17.3769,10.2916,23.2056,20.3746]
Tumor	[21.8991,18.1287,9.0217,15.4336,44.8240,58.2788]

Diagnostic result using the above patterns on the test data set

100% for normal, 92.30% for tumor, 96.10% for normal and tumor, 90.20% for crossvalidation

Diagnostic result using the above patterns on an independent data set from another laboratory

100% for normal, 96.00% for tumor, 97.10% for normal and tumor

model 3 (four genes)

Gene	GenBank no.	Gene name
9850	M84526	Human adipsin/complement factor D mRNA, complete cds
6715	N36638	Unknown
10875	L78132	Human prostate carcinoma tumor antigen (PCTA-1) mRNA, complete cds
6185	X07732	Human hepatoma mRNA for serine protease hepsin

Diagnostic expression patterns for the above four genes

Normal	[84.0281,15.5463,8.9924,27.3297]
Tumor	[23.7969,8.0311,11.8138,59.2811]

Diagnostic result using the above patterns on the test data set

100% for normal, 100% for tumor, 100% for normal and tumor, 90.20% for crossvalidation

Diagnostic result using the above patterns on an independent data set from another laboratory

100% for normal, 88.00% for tumor, 91.20% for normal and tumor

The ICA-based algorithm (see text for details) was sequentially performed three times on prostate cancer data sets, and the expression profiles of 12 600 genes using Affymetrix U95Av2 arrays in 50 normal and 52 prostate cancer tissue samples and¹¹ three diagnostic models (models 1–3) were obtained.

$$V_{\text{test}} = \begin{pmatrix} 223 \\ 298 \\ 123 \\ 39 \\ 101 \\ 142 \end{pmatrix}$$

(b) Calculating its Euclidean distances to two diagnostic expression patterns on the six genes of interest (Table 2)

$$V_{\text{normal}} = \begin{pmatrix} 84.3491 \\ 50.3947 \\ 17.3769 \\ 10.2916 \\ 23.2056 \\ 20.3746 \end{pmatrix}, \quad V_{\text{tumor}} = \begin{pmatrix} 21.8991 \\ 18.1287 \\ 9.0217 \\ 15.4336 \\ 44.8240 \\ 58.2788 \end{pmatrix}$$

we obtained the distance between V_{test} and V_{normal} $d1 = 336$ and the distance between V_{test} and V_{tumor} $d2 = 377$.

(c) Obviously, $d2 > d1$, so we can diagnose the No. 4 sample as a normal sample, as expected.

The diagnostic results using three models on the independent data set are summarized in Table 2 (the diagnostic data are available in Supplementary Information). The results reveal that our three models also perform well with high accuracy (88.2% for model 1, 97.1% for model 2 and 91.2% for model 3).

Finally, we applied the present three diagnostic models to diagnose 10 nonrecurrent and eight recurrent prostate cancer samples. The result demonstrates that our three models also have very good performance: 88.9% for model 1, 83.3% for model 2 and 88.9% for model 3 (the diagnostic data are available in Supplementary Information). Thus, our diagnostic model is capable of diagnosing a tumor from unknown prostate samples including nonrecurrent or recurrent prostate cancer with excellent accuracy.

However, it remains possible that these apparently good models were obtained from the data set by chance. We therefore asked what could be the probability for successfully obtaining the diagnostic models presented above by chance alone? To explore this issue, we ran the program 1000 times, based on 1000 permutations of normal and tumor sample tags. For each permutation, we tried to find a diagnostic model using the same number of genes (4–6 genes for prostate cancer and 4–10 genes for colon cancer). The outcome of this control study was that 21 and 16 of the 1000 permutations generated models with $>90\%$ crossvalidation accuracy for prostate and colon cancers, respectively. This shows that the probability for obtaining the above diagnostic models by chance alone was of the order of 1.6–2.1%.

Discussion

The very large amount of gene expression information and noisy data provided by microarray technology leads to

difficulties in both basic research and clinical applications. ICA may be an ideal technique for reducing the dimension of data and for separating the experimental noise from expression data.⁸ The biomarkers selected by our ICA-based method should be examined for their roles in disease etiology. A total of 12 and nine different genes were identified for colon cancer (Table 1) and prostate cancer (Table 2), respectively. Among these genes, the majority of colon cancer genes were identified and discussed in previous studies,^{13–15} except two genes, Integrin Alpha 6 and Wee1 Hu genes, which are newly identified by the present ICA study. In contrast, for prostate cancer, only one gene, Hepsin, was identified and discussed by Welsh *et al*;¹² the remaining eight genes are newly identified by the present ICA study: Adipsin/complement factor D, KET protein, Insulin-Like growth factor 2, Elongation factor 1 alpha-2, PCTA-1, Human beta-migrating plasminogen activator inhibitor 1, Alpha-methylacyl-CoA racemase (AMACR) and an unknown gene from human melanocyte. These genes have been implicated in human cancers (see Supplementary Information).

It should be emphasized that the genes identified by ICA may not be genes with significantly differential expression between normal and tumor samples. For example, while it has already been investigated as a possible cause for predisposition to prostate cancer but without analysis of its regulation,¹⁶ the average expression level of PCTA-1 in tumor samples is about 1.5-fold its average expression level in normal samples. Naturally, such a gene will be eliminated if using the conventional two-fold selection criteria. While ICA identified this gene as a biomarker, and model 3 including this gene can achieve 100% classification accuracy for prostate cancer (Table 2). In fact, PCTA-1 encodes a member of the galectin family. The galectins have been implicated in many essential functions including development, differentiation, cell–cell adhesion, cell–matrix interaction, growth regulation, apoptosis and RNA splicing. PCTA-1 has been considered as a surface marker associated with prostate cancer.¹⁷

Here, we obtained three models for both colon and prostate cancer. However, other models reached similar classification accuracy, as Xiong *et al*¹⁸ indicated that the optimal or near-optimal sets of genes for classifying tumor and normal tissues are not unique. This just reflects the complexity of microarray data. We expected, as seen in Table 1, that the more the genes were included in the model, the better the accuracy of the classification was. In contrast, however, Table 2 displays another situation where a model with fewer genes could get better accuracy (model 3 vs model 2). This prompted us to make further tests: interestingly, we found that as more genes are included in the model, the classification accuracy is not systematically increased as expected. Similar results were found in previous reports.^{14,19} This highlights the conclusion that no correlation between the number of genes and the

Table 3 Comparison of ICA-based algorithm with other techniques used on the same colon cancer data set

Methods	Classification accuracy (%)	Diagnostic pattern available for other data	Reference
Genetic algorithm (GA) and K-nearest neighbor method	75	No	14
Two-way clustering	87.10	No	10
Principle component analysis (PCA)	87.10	No	19
Hierarchical clustering	90.30	No	21
Fisher's linear discriminant analysis	92	No	18
Partial least squares (PLS)	93.50	No	19
Correspondence analysis	93.50	No	13
Pairwise gene expression ratio	93.60	No	22
Recursive partitioning method	98	No	15
ICA	92	Yes	This study

classification accuracy exists. This can be accounted for by the fact that not all genes are equally relevant to the normal vs tumor sample discrimination, while the inclusion of more genes introduces more high-dimensional noise, hence decreasing the classification accuracy. As a matter of fact, it is a well-known statistical property that too many feature variables (genes) can harm a discriminator's performance.²⁰ This suggests that we should favor methods that try to extract as few genes as possible as biomarkers. The present work is an attempt in this direction.

Among the prostate cancer patients, up to 30% of them undergoing radical prostatectomy will relapse. The challenge is therefore to identify patients at risk for relapse so that relevant targeted therapeutics could be focused on that particular subpopulation. In the present study, by ICA we identified a few new biomarkers of the disease, and using their expression patterns we could achieve good diagnosis for the nonrecurrent/recurrent data set. This should be helpful for finding robust prognostic markers that are capable of identifying patients at risk of relapse following local therapy.

Finally, we compared other techniques using the same colon cancer data set. Table 3 shows that the present ICA-based algorithm offers a significant advantage over the other methods, which classified two subgroups (tumor vs normal) but do not provide a specific expression pattern of few genes for diagnosis. In contrast, our method not only achieves good classification accuracy, but, because it uses ICA, explicitly extracts a specific expression pattern, thus allowing one to perform diagnosis for any blind sample through a small microarray of few genes. As seen above, for prostate cancer, we have used three different diagnostic expression patterns, obtained from the first laboratory data set, to directly diagnose an independent data set from another laboratory. The performance of the diagnosis was remarkably consistent. However, we need to explore more data from different laboratories to validate these diagnostic models. In conclusion, diagnostic patterns obtained from as many clinical samples as possible using our ICA-based method should be undertaken, in order to clinically

provide quantitative diagnostic information for an individual cancer patient.

Acknowledgements

We thank the Hong Kong Innovation and Technology Fund BIOSUP-PORT program for supporting the present research.

References

- Ramaswamy S, Tamayo P, Rifkin R *et al*: Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA* 2001; **98**: 15149–15154.
- Simon R, Radmacher MD, Dobbin K, McShane LM: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003; **95**: 14–18.
- Hyvarinen A, Oja E: Independent component analysis: algorithm and applications. *Neural Netw* 2000; **13**: 411–430.
- Stone JP, Porrill J, Porter NR, Wilkinson ID: Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions. *Neuroimage* 2002; **15**: 407–421.
- Vigario R, Sarela J, Jousmaki V, Hamalainen M, Oja E: Independent component approach to the analysis of EEG and MEG recordings. *IEEE Trans Biomed Eng* 2000; **47**: 589–593.
- Vigario RN: Extraction of ocular artefacts from EEG using independent component analysis. *Electroencephalogr Clin Neurophysiol* 1997; **103**: 395–404.
- Liebermeister W: Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 2002; **18**: 51–60.
- Lee SI, Batzoglou S: Application of independent component analysis to microarrays. *Genome Biol* 2003; **4**: R76.
- Saidi SA, Holland CM, Kreil DP *et al*: Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene* 2004; **23**: 6677–6683.
- Alon U, Barkai N, Notterman DA *et al*: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 1999; **96**: 6745–6750.
- Singh D, Febbo PG, Ross K *et al*: Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002; **1**: 203–209.
- Welsh JB, Sapinoso LM, Su AI *et al*: Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* 2001; **61**: 5974–5978.
- Kishino H, Waddell PJ: Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Inform Ser Workshop Genome Inform* 2000; **11**: 83–95.
- Li L, Darden TA, Weinberg CR, Levine AJ, Pedersen LG: Gene assessment and sample classification for gene expression data

- using a genetic algorithm/k-nearest neighbor method. *Comb Chem High Throughput Screen* 2001; **4**: 727–739.
- 15 Zhang H, Yu CY, Singer B, Xiong M: Recursive partitioning for tumor classification with gene expression microarray data. *Proc Natl Acad Sci USA* 2001; **98**: 6730–6735.
- 16 Maier C, Rosch K, Herkommer K *et al*: A candidate gene approach within the susceptibility region PCaP on 1q42.2–43 excludes deleterious mutations of the PCTA-1 gene to be responsible for hereditary prostate cancer. *Eur Urol* 2002; **42**: 301–307.
- 17 Gopalkrishnan RV, Roberts T, Tuli S, Kang D, Christiansen KA, Fisher PB: Molecular characterization of prostate carcinoma tumor antigen-1, PCTA-1, a human galectin-8 related gene. *Oncogene* 2000; **19**: 4405–4416.
- 18 Xiong M, Li W, Zhao J, Jin L, Boerwinkle E: Feature (gene) selection in gene expression-based tumor classification. *Mol Genet Metab* 2001; **73**: 239–247.
- 19 Nguyen DV, Rocke DM: Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 2002; **18**: 39–50.
- 20 McLachlan GJ: *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley, 1992.
- 21 Li W, Fan M, Xiong M: SamCluster: an integrated scheme for automatic discovery of sample classes using gene expression profile. *Bioinformatics* 2003; **19**: 811–817.
- 22 Yap Y, Zhang X, Ling M, Wang X, Wong Y, Danchin A: Classification between normal and tumor tissues based on the pair-wise gene expression ratio. *BMC Cancer* 2004; **4**: 72–88.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)