

---

# Dimension Reduction of Microarray Data with Penalized Independent Component Analysis

---

Han Liu   Rafal Kustra\* \*

## Abstract

In this paper we propose to use ICA as a dimension reduction technique for microarray data. All microarray studies present a dimensionality challenge to the researcher: the number of dimensions (genes/spots on the microarray) is many times larger than the number of samples, or arrays. Any subsequent analysis must deal with this dimensionality problem by either reducing the dimension of the data, or by incorporating some assumptions in the model that effectively regularize the solution. In this paper we propose to use the ICA approach with a regularized whitening technique to reduce the dimension to a small set of independent sources or latent variables, which then can be used in downstream analysis. The elements of the mixing matrix can themselves be investigated to gain more understanding about the genetic underpinnings of the process that generated the data. While a number of researchers have proposed ICA as a model for the microarray data, this paper is different in an important aspect: we focus on ICA as a dimension reduction step which leads us to the generative model formulation that applies the ICA in an opposite way to most other proposals in this field.

## 1 Introduction

DNA microarray experiments generate large datasets with expression values for thousands of genes but no more than a few dozens of samples. Each microarray sample  $i$  can be represented as  $x_i^T = (x_{i1}, \dots, x_{ip})$ , and we collect all microarrays in a  $p \times n$  matrix  $X$ .

A typical situation where the large number of genes compared with small sample size presents a challenge has been explored in [13]. In this paper, three well-known microarray datasets (AML/ALL Leukemia, Lymphoma, and NCI-60) have been used to compare different classification techniques. The common feature of these datasets is that a few dozen tissues from different types of cancers have been

---

\*Rafal Kustra is an assistant professor of Department of Public Health, Biostatistics division, from University of Toronto. Han Liu is a graduate student of Department of Computer Science from University of Toronto

microarrayed to obtain expression profiles of a few thousand genes. It is of interest to build a classifier that would be able to reliably distinguish among different types of cancers based solely on the expression profile. This by itself could provide significant help in diagnosing and staging cancers, and the classification/discrimination rules built by these classifiers could be examined and interpreted to gain more insight into the molecular working of different cancers.

In most of the previous studies, including [13], univariate gene selection has been used extensively for reducing the number of genes to be considered before classification. Several selection criteria have been used in the literature, e.g. the BSS/WSS ratio (equivalent to F statistic) used in [13], the  $t$  statistic [1], the Wilcoxon's rank sum statistic [2] or Ben Dor's combinatoric "TNoM" score [3]. The main advantages of gene selection are its simplicity and interpretability. However proper gene selection based on univariate statistics is a difficult problem akin to the differential gene expression analysis and its associated multiple testing challenge, both the topics of much current research (e.g. [16] [15] [14]). As is already evident from these and other papers, some amount of information sharing among genes is essential even when one seeks to select individual genes as the final goal.

Another approach to dimension reduction is via multivariate methods, which have a flavor of feature extraction, and naturally consider some aspects of joint distribution of all genes. Unlike gene selection, the general idea is to construct a limited set of  $k$  components  $z_1, \dots, z_k$  which are functions of the original variables. In this paper we concentrate on linear functions. Thus we search for vectors  $b_j$  to construct features  $z_j = b_j^T \mathbf{x}$  for  $1 \leq j \leq k$ . (In all that follows we assume that the data has been centered). Let  $\mathbf{Z} = (z_1, \dots, z_k)^T$  be the  $k \times n$  matrix having the components in its rows. The most well known linear features are usually constructed by a PCA method (also known as Karhunen-Loeve expansion in pattern recognition, and typically implemented with Singular Value decomposition) seeks components that are uncorrelated and have maximum variance [4]:

$$E(\mathbf{z}\mathbf{z}^T) = \mathbf{D}_k, \quad (1)$$

where  $\mathbf{D}_k$  stands for a diagonal matrix of rank  $k$ . These components (typically scaled by inverse roots of elements of  $\mathbf{D}_k$  to obtain unit-variance) will then serve as new predictor variables in the further analysis.

In this paper, we propose a new dimension reduction technique for specific genomics applications based on Independent Component Analysis (ICA) [5]. Unlike PCA, which only considers variance, ICA is able to exploit higher order statistics which may contain significant complementary information. This is particularly important when the distribution of data differs significantly from normal, which for at least two reasons is often the case with microarray expression data. One reason is that the gene expression data is often heavy-tailed [6], since great majority of genes do not react strongly to different experimental conditions, while the ones that do, have very strong signals. Another reason is that due to very limited sample sizes, the Central Limit Theorem often does not apply in these problems.

Our approach models logarithms of expression profile of specific genes across different samples to reduce the microarray data to a small set of independent sources or latent variables, which then can be used in place of genes in subsequent analysis. The elements of the mixing matrix can themselves be investigated to gain more understanding about the genetic underpinnings of the process that generated the data. This method provides a biologically meaningful approach to dimension reduction and has more promising statistical properties than traditional methods now used.

## 2 Methodology

### 2.1 General Framework

The relative expression levels of  $p$  genes of a sample organism are probed simultaneously by a single microarray. A series of  $n$  arrays, which physically identical, probe the genome-wide expression levels in  $n$  different samples. Let the  $p \times n$  matrix  $\mathbf{X}$  denote the full expression data. For cDNA studies, the elements of  $\mathbf{X}$  would be log-ratio of signal-to-reference channels, while for oligonucleotide arrays these would be log-PM measurements, possibly offset by the MM measurements and aggregated in a suitable way over all probes of a gene. We also assume that a suitable normalization was performed on the microarrays, so that columns of  $\mathbf{X}$ , which correspond to arrays in the data, represent samples from the same family of distributions.

By viewing the expression profile of a single microarray as a realization of a random vector, we model it as a mixture of small number of latent components which are statistically independent. Mathematically, suppose that each gene on a microarray is a different linear combination of the same  $k$  independent random variables,  $\mathbf{s} = (s_1, \dots, s_k)^T$ , which are termed Independent Components. Biologically, these components may correspond to distinct biological processes that are affected by the experimental design, and in terms modulate the expression level of particular genes through the various genetic regulatory mechanisms such as transcription factors.<sup>1</sup>

We can express this model consistently in the generative form of ICA.

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ a_{21} & \dots & a_{2k} \\ \vdots & \vdots & \vdots \\ a_{p1} & \dots & a_{pk} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{pmatrix} \quad (2)$$

which for a particular dataset can be put in a matrix form:

$$\mathbf{X}_{p \times n} = \mathbf{A}_{p \times k} \mathbf{S}_{k \times n}$$

Each of the vector  $s_1, \dots, s_k$  can be viewed as an independent random source which may have some biological meaning. Then, ICA can be applied to find a matrix  $\mathbf{W}$  that provides the transformation  $\mathbf{Y} = (y_1, \dots, y_k)^T = \mathbf{W}^T \mathbf{X}$  of the observed matrix  $\mathbf{X}$  under which the transformed random variables  $y_1, \dots, y_k$  called the Estimated Independent Components, are as independent as possible [7]. Under certain mathematical conditions (see also section 2.2), the estimated random variables  $y_1, \dots, y_k$  are close approximations of  $s_1, \dots, s_k$ , up to permutation and scaling. In matrix notation

$$\mathbf{Y}_{k \times n} = \mathbf{W}_{k \times p}^T \mathbf{X}_{p \times n} \quad (3)$$

From equation 3, we see that the data are mapped from a  $p \times n$  space to a reduced  $k \times n$  space, and when  $k \ll p$ , the dimension is significantly reduced. In the new space, the data are represented by the matrix  $\mathbf{Y}$ , the  $k$  rows of this matrix are  $k$  independent predictors. A typically hard problem, akin to determining a valid number of principal components or number of clusters, is the selection of  $k$ . For the purposes of this paper we just investigate three levels of  $k$ : high,  $k \approx n/2$ ; low:  $k \approx 3 \sim 5$ ; and medium level which is a compromise between those two.

---

<sup>1</sup>Two caveats: we do not necessarily imply causality here, merely postulate a model that may explain comparability in a simplified way. Also, in practice, given a small number of data (arrays), each component may be a mixture of underlying processes with perhaps one or two dominating the rest in the context of particular experimental setup and intervention factors

## 2.2 Regularized Whitening Procedure

Most ICA algorithms, require a “whitened” data, with an identity covariance matrix. This is more than a numerical requirement, in that ICA is attempting to decompose the data beyond the first two moments which is the goal of PCA and related methods. Typically, second-moment would dominate larger moments and be detrimental to the ICA goal.

Many whitening methods are based on *eigenvalue decomposition (EVD)* as shown in [7]. In which, the centered data  $\mathbf{x}$  are transformed into  $\tilde{\mathbf{x}}$ , such that  $E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\} = I_p$ . With eigenvalue decomposition of the estimated covariance matrix,  $1/n\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T$ , normally used whitening transformation is  $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}^{-\frac{1}{2}}\mathbf{X}$ . However, when  $n \ll p$ , the eigenvalue matrix  $\mathbf{D}$  is far from full rank, thus the inverse of  $\mathbf{D}$  does not exist.

We propose to use a penalized estimates of the covariance matrix to whiten the data. In general, penalized estimates are biased version of the un-penalized ones, but exhibit lower variance, if the penalty is chosen a’priori. This can lead to estimates which have a lower mean-square error, especially in low S/N or high-dimensional settings.

For the purposes of this paper, we propose a very simple penalization scheme based on the identity matrix which, however, has been proven effective in high-dimensional settings [17]. If we use:

$$\tilde{\Sigma} = 1/n\mathbf{X}\mathbf{X}^T + \lambda I_p$$

as a penalized estimate of the gene-gene covariance matrix, the whitening procedure becomes:

$$\tilde{\mathbf{X}} = \mathbf{U}(\mathbf{D} + \lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{X} \quad (4)$$

In equation 4, when the  $\lambda$  that was chosen as 75 quartile of all non zero eigenvalues  $\lambda_1, \dots, \lambda_{n-1}$  of the covariance matrix. In more automatic procedure, one can use leave-one-out cross validation to tune the regularization factor  $\lambda$ .

## 2.3 Independent Component Analysis Algorithm

We denote  $\mathbf{X}$  the corrected logarithms of expression profile after standardization, and start from a model of the form  $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$ , where the  $\mathbf{Y} = (y_1, \dots, y_k)$  are independent sources and  $\mathbf{W}^T$  is the unmixing matrix. The ICA algorithm we adopt is called **FastICA** which searches the corresponding  $\mathbf{W}^T$  by minimizing the mutual information as follows:

$$I(y_1, \dots, y_k) = \sum_{i=1}^k H(y_i) - H(\mathbf{X}) + \log |\det(\mathbf{W}^T)| \quad (5)$$

where  $H(y)$  represents the entropy for random variable  $y$  with density  $f(y)$  and defined as  $H(y) = -\int f(y) \log f(y) dy$ . After this step, the data have already been mapped into a new feature space and when  $k \ll p$ , the dimension is reduced greatly. Also, we can estimate the mixing matrix  $\mathbf{A} \approx \mathbf{W}^T$ .

As a result, the ICA method yields latent sources which are statistically independent. There are mainly two aspects we are of great interest: Given the generative model  $\mathbf{X}_{p \times n} = \mathbf{A}_{p \times k}\mathbf{S}_{k \times n}$ , the mixing matrix  $\mathbf{A}$  is of great interest to analysis. For a specific gene, one of the elements  $a_{ij}$  ( $1 \leq i \leq p$  and  $1 \leq j \leq k$ ) represents the effect of the  $j$ th latent source on the  $i$ th gene across  $n$  different conditions (arrays). If the generative model does hold, based on this information, we can predicate whether

this latent factor is (positive or negative) "active" under the conditions. The other aspect is when fixing a specific latent factor, the distribution of the elements of matrix  $\mathbf{A}$  could be a good indication for analyzing the behavior of specific genes in different latent factors. Given a threshold, the distribution of gene expression profile in a given latent factor generally features a small number of significantly over-expressed or under-expressed genes, which kind of "dominate" this latent factor.

When the algorithm was implemented, the approximations developed in [18], based on the maximum entropy principle. we get

$$H(y_{gauss}) - H(y) \approx c[E\{G(y)\} - E\{G(v)\}]^2 \quad (6)$$

where  $G$  is practically any non-quadratic function,  $c$  is an irrelevant constant, and  $v$  is a standardized Gaussian variable. The choice of  $G$  will be discussed later in section 3.

## 2.4 Comparison with Related Works

Some other researchers also applied ICA for microarray analysis. Liebermeister [8] and Chiappetta [9] first proposed using linear ICA for microarray analysis to extract expression modes, where each mode represents a linear influence of a hidden cellular variable. Su-In Lee [10] gave out a systematic analysis of the applicability of ICA as an analysis tool in diverse datasets. Given a  $p \times n$  microarray expression profile matrix  $\mathbf{X}$ , they assume a generative model  $x_j^{(i)} = \sum_{\nu} a_{i\nu} s_{\nu}$ , where  $j$  indexes genes and superscript  $i$  indexes arrays. This uses separate realization of the random component vector,  $\mathbf{s}$ , for each gene,  $j$ , rather than for each array as we propose. We believe that our application of ICA to genomic data better expresses the biological reality of co-expression and leads to better modelling of comparability among genes. We also believe it is closer in spirit to the original motivation of ICA that came from Blind Source Separation community: in our settings the spots on the array act as microphones and the underlying biological processes are the independent speakers. It is also true that our application enables us to perform the dimension reduction using ICA, which is the goal of this paper.

## 3 Experiments

To evaluate the performance of our method, the ICA based method has been applied to three real world datasets: NCI 60, Leukemia and Yeasts data.

### 3.1 Dataset

**NCI 60:** In this dataset, cDNA were used to examine the variation in gene expression among the 60 cell lines from the National Cancer Institute's anticancer drug screen known as NCI 60 dataset [11]. The 60 cell lines include: 7 breast, 6 central nervous system(CNS), 7 colon, 6 leukemia, 8 melanoma, 9 nonsmall-cell lung carcinoma (NSCLC), 6 ovarian, 2 prostate, 8 renal, and 1 unknown (ADR-RES). Gene expression was studied using microarrays with 9,703 spotted cDNA sequences. In each hybridization, fluorescent cDNA targets were prepared from a cell line mRNA sample (fluorescent dye Cy5) and a reference mRNA sample obtained by pooling equal mixtures of mRNA from 12 of the cell lines (fluorescent dye Cy3). For our experiment, we make classification for 8 classes (the two prostate cell line observations were excluded out because of their small class size). After screening out genes with missing data points, the data are collected into a  $5244 \times 61$  matrix  $\mathbf{X} = (x_{ij})$ ,

where  $x_{ij}$  denotes the logarithmic of the Cy5/Cy3 fluorescence ration for gene  $i$  in mRNA sample  $j$ . Also, the standardization of the data have been performed as described above, 43 samples are for the training dataset while 18 samples as testing dataset.

**Leukemia:** The data here is the acute leukemia data set published by Golub et al. [12]. The training dataset consisted of 48 bone marrow samples with 27 ALL and 11 AML (from adult patients). The test dataset consisted of 24 bone marrow samples as well as 10 peripheral blood specimens from adults and children (20 ALL and 14 AML). Four AML samples from the independent dataset were from adult patients. The gene expression intensities were obtained from Affymetrix high density oligonucleotide microarrays contacting probes for 6,816 genes. The preprocessing procedure is the same as above.

**Lymphonma:** This dataset contains gene expression levels of 3 class most prevalent adult lymphoid malignancies: 42 samples of diffuse large B-cell lymphoma, 9 observations of follicular lymphoma and 11 cases of chronic lymphocytic leukemia. the total sample size is  $n = 62$ , and the expression of  $p = 4,026$  well-measured genes, preferentially expressed in lymphoid cells or with known immunological or oncological importance is documented. More information on these data can be found in Alizadeh et al. [19], the imputation of missing values and standardizations of the data is expressed in [13].

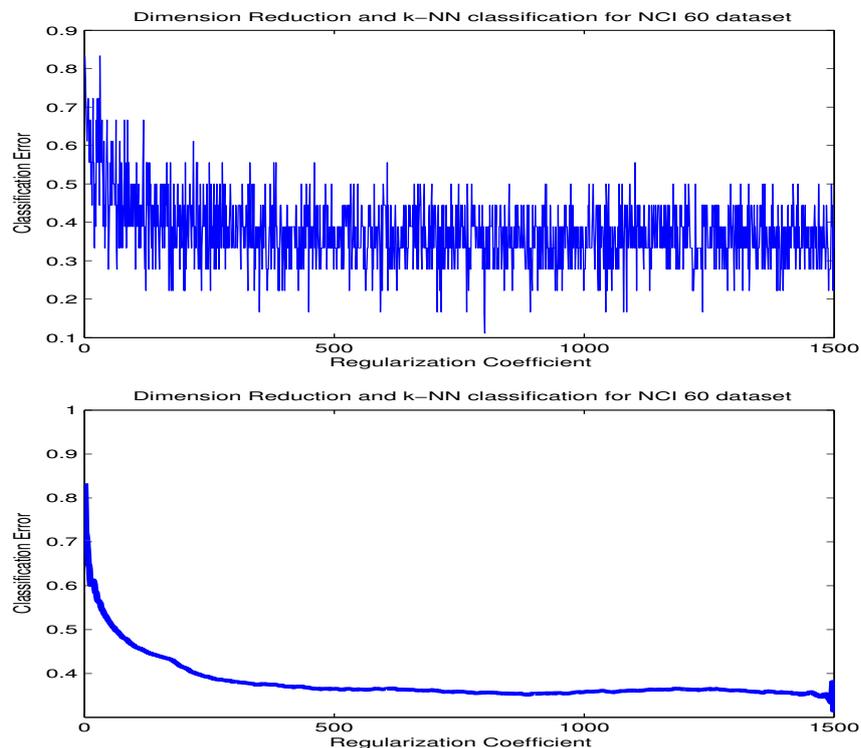


Figure 1: Illustration of the regularization whitening to the prediction performance on NCI60 dataset with k-NN classifier, The upper figure shows the misclassification error as the regularization coefficient ranging from 1 to 1500, the lower figure shows the smoothed version

### 3.2 Study Design and Experimental Results

After dimension reduction with ICA, we apply discriminant analysis and k Nearest Neighbor methods for classification. The results are promising. Practically, some common used contrast functions are shown as below

$$\begin{aligned} G_1(u) &= \frac{1}{a_1} \log \cosh(a_1 u) \\ G_2(u) &= -\frac{1}{a_2} \exp(-a_2 u^2/2) . \\ G_3(u) &= \frac{1}{4} u^4 \end{aligned} \quad (7)$$

where  $1 \leq a_1 \leq 2$ ,  $a_2 \approx 1$  are constants, and piecewise linear approximations of  $G_1$  and  $G_2$  are also used. However, in the experiment we found that when choosing  $G = \frac{1}{3} u^3$ , the performance is much better than above choices.

Normally,  $\lambda$  can be chosen as the 75 quartile of all the nonzero eigenvalues. However, here, to visualize the effect of regularization coefficient  $\lambda$  on the prediction performance, we increased  $\lambda$  as an integer from 0 to 1,500. For each  $\lambda$ , we reduced the dimension of the data from  $n$  to  $k$ . For the three levels of  $k$ , we found that when  $k = 20$ , the 1 Nearest Neighbor classifier performed the best. The experimental result of 1 Nearest Neighbor classifier on the NCI 60 dataset was shown in figure 1. It can be seen that without regularization, or regularized with very small  $\lambda$ , the misclassification is extremely high; in fact, it is a sign of overfitting on the covariance matrix. As the value of  $\lambda$  increased, the regularization identity matrix tended to dominate and the prediction performance tended to be stabilized. For more details, please refer to [17]. Similar experiments were also conducted on the other datasets with different discriminant methods, including DLDA, DQDA and k-NN. The prediction performance is significantly better than the results shown in [13], and the distribution of the unmixing matrix shows strong patterns. which sheds a light of the promise of this ICA based dimension reduction paradigm.

## 4 Conclusion and Future Work

In this extended abstract we propose to use ICA as a dimension reduction technique for microarray analysis. Which could be viewed as an extension for PCA based method. With a regularized whitening technique, we reduce the dimension to a small set of independent sources or latent variables, which then can be used in subsequent discriminant analysis. The elements of the mixing matrix can themselves be investigated to gain more understanding about the genetic underpinnings of the process that generated the data. While a number of researchers have proposed ICA as a model for the microarray data, this paper is different in an important aspect: we focus on ICA as a dimension reduction step which leads us to the generative model formulation that applies the ICA in an opposite way to most other proposals in this field. The significant improvement of the prediction performance also illustrate that it is a promising method.

In this purpose, the ICA algorithm we used is a general purpose blind source separation algorithm, even though we have shown that it can be used as a effective and efficient dimension reduction technique for gene expression data, it's not specially tailored for dimension reduction task. We believe that a dimension reduction driven ICA algorithm may worth some further investigation.

## References

- [1] Hedenfalk,I., Duggan,D. etc. (2001) Gene expression profiles in hereditary breast cancer. *N Engl J Med* 335, pp. 539-548.

- [2] Dettling, M., & Buhlmann, P (1995) Boosting for tumor classification with gene expression data. *Bioinformatics* 19, pp. 1061-1069.
- [3] Ben-Dor, A., & Brijm, L., Frideman, N., Nachman, I., Schummer, M. Yakhini, Z., (2000) Tissue classification with gene expression profiles., *Journal of Computational Biology* 7, pp. 559-584.
- [4] Alter O, Brown PO, & Botstein D., (2000) Singular value decomposition for genome-wide expression data processing and modeling, *Proc Natl Acad Sci USA* 97, pp.10101-10106.
- [5] Hyvarinen, A. and Oja, E. (1999) Independent component analysis: algorithm and applications, *Neural networks* 13, pp.411-430.
- [6] Schena, M., Shalon, D. Davis, R.W. and Brown, P.O., (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* 270, pp.467-470.
- [7] Hyvarinen A., (1999) A survey of independent component analysis, *Neural Computing Surveys* 2, pp.94-128.
- [8] Liebermeister, W., (2002) Linear modes of gene expression determined by independent component analysis, *Bioinformatics* 18, pp.51-60.
- [9] Chippetta, P., Roubaud, M. and Torresani, B., (2002) Blind source separation and the analysis of microarray data, *Proc of JOBIM'02*, pp.131-136.
- [10] Su-In, L., Serafim, B., (2003) Application of independent component analysis to microarrays, *Genome Biology* 4, R76.
- [11] Ross, D.T., Scherf, U., etc. (2000) Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics* 24, pp.227-234.
- [12] Golub, T.R., Slonim, D.K., etc. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, pp.531-537.
- [13] Dudoit, S., J. Fridlyand, and T.P. Speed Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97(457), 77-87.
- [14] Tusher VG, Tibshirani R, Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA.*, 98(9) pp.5116-21.
- [15] Efron B, Tibshirani R. (2002) Empirical bayes methods and false discovery rates for microarrays *Genet Epidemiol.*(23) pp.70-86.
- [16] Kendziora CM, Newton MA, Lan H, Gould MN. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles *Stat Med.* 22(24) pp.3899-914.
- [17] Rafal Kustra, Stephen C. Strother. (2001) Penalized Discriminant Analysis of [15O]-water PET Brain Images with Prediction Error Selection of Smoothness and Regularization. *IEEE Trans. Med. Imaging* 20(5) pp.376-387.
- [18] Hyvarinen, A. (1998) New approximations of differential entropy for independent component analysis and projection pursuit. *In Advances in Neural Information Processing Systems* 10 pp.273-279. MIT press.
- [19] Alizadeh, A., Eisen, M., Davis, R., etc. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403 pp.503-511.