

**TO PREDICT DISEASE OUTCOME: CLINICAL RISK FACTORS
PLUS GENETIC-STAGING FOR CANCER.**

Shu-Kay Ng

School of Medicine, Logan Campus, Griffith University, Meadowbrook, Australia
Email: s.ng@griffith.edu.au

ABSTRACT

Studies on genetic profiling demonstrate its potential utility for classifying tumours, leading to a “genetic-staging” system for predicting disease outcomes. A precise prediction of individual disease outcome is important to identify patients who have a high risk of disease recurrence, and to tailor treatments to the individual patient. Genes, however, are not the sole determinants of disease outcomes. Non-genetic factors have roles in many stages of tumourigenesis, and the simultaneous use of genetic-staging and clinical risk factors may therefore improve the prediction of disease outcome. In this paper, we aim to quantify the prognostic value of genetic-staging from gene expressions by using mixture model-based clustering methods. We also investigate via the use of logistic regression whether a more accurate prediction of disease outcome can be obtained by using genetic-staging in conjunction with clinical risk factors. The proposed method is illustrated using a real example of breast cancer data. It shows that genetic-staging provides significant additional prognostic information when it is obtained by applying sophisticated model-based clustering method for the identification of marker-genes that are relevant to predict disease outcomes.

KEYWORDS

Clustering method, EMMIX-GENE, EMMIX-WIRE, Logistic regression, Mixture model, Pattern recognition.

2000 Mathematics Subject Classification: Place Classification here. Leave as is, if there is no classification.

1 INTRODUCTION

Treatment for individual patients is chosen according to various criteria, such as the clinico-pathologic factors of the disease and the demographic characteristics of the patients (Caldas and Aparicio, 2002). A more accurate prediction of individual disease outcome is of paramount importance to identify high-risk patients who may require carefully tailored health care or further therapy for possible recurrence of the disease after the initial treatment (Brand, Brand, and den Baumen, 2008). For example, patients may be separated by disease outcomes as good- and poor-prognosis groups. A precise prediction of prognosis outcome will help to guide the treatment plan and health care for individual patients. Studies on the analysis of gene expression in colon, breast, and other tumours demonstrate the potential utility of expression signatures for classifying patients (Lapointe and Li, 2004; van de Vijver et al., 2002; Xiong et al., 2001). However, genes are not the sole determinants of disease outcomes (Carlsten and Burke, 2006; Rhodes and Pollock, 2006). Environmental and other non-genetic risk factors have roles in many stages of tumourigenesis, which may lead to disparity in patient clinical characteristics (Halliday et al., 2004). The simultaneous use of gene expression signatures and clinical data can improve the prediction of disease outcome in situations where the clinical data contains information beyond that provided by the genetic microarrays; see Ben-Tovim Jones et al. (2005).

In this paper, we aim to quantify the prognostic information from gene expression signatures using a “genetic-staging” for predicting disease outcome of cancer patients. We also investigate if a more accurate prediction of disease outcome can be achieved by using genetic-staging for cancer in conjunction with clinical risk factors.

For the formation of genetic-staging for cancer, the idea is to classify patients, based on their gene expression signatures, into different groups that correspond to different genetic stages for the disease. However, the classification of patients into different genetic stages is a nonstandard problem in parametric statistical analysis. It is because typical microarrays contain expression data on thousands of genes, which is much greater than the number of tissue samples from patients (typically between 10 and 100). Multivariate approaches such as principal component analysis and partial least squares have been proposed to reduce the dimension of gene signatures for classifying cancer patients (Nguyen and Rocke, 2002; West et al., 2001). However, as described in McLachlan, Do, and Ambrose (2004), practical problems arise because biological interpretation of the principal components obtained using these approaches is not straightforward. Alternatively, combinatorial searches through the possible subsets of genes have been considered (Kudo and Sklansky, 2000).

These heuristic search algorithms are, however, computationally too prohibitive to search for an optimal or near optimal subset of genes (McLachlan, Do, Ambroise, 2004). More recently, increasing attention on this high-dimensional problem is being given to advanced methodologies on model-based clustering (Kim, Tadesse, and Vannucci, 2006; McLachlan, Do, and Ambroise, 2004). In this paper, we consider two mixture model-based clustering methods to handle this dimensionality problem. Prediction of disease outcomes will be finalized by classifying patients on the basis of their genetic-staging as well as corresponding clinical risk factors, via the logistic regression model. Figure 1 outlines the proposed method for the prediction of disease outcomes.

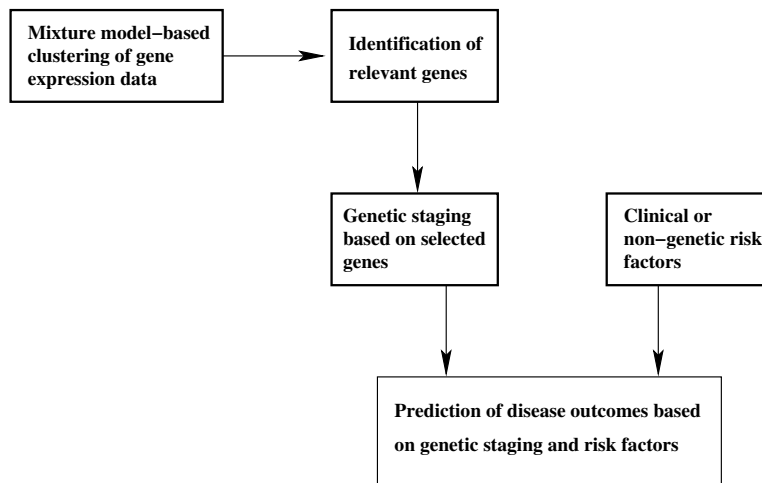


Figure 1: Proposed method for the prediction of disease outcomes

The rest of the paper is organized as follows: Section 2 presents the methodologies for the formation of genetic-staging based on model-based clustering of gene expressions data. We describe in Section 3 the prediction of individual disease outcomes by combining the genetic-staging information with patient clinical characteristics. In Section 4, the method is applied to the breast cancer dataset of van't Veer et al. (2002) that consists of 5000 gene expression profiles and 6 binary clinical variables from 78 patients. Section 5 ends the paper by presenting some concluding remarks and discussion.

2 GENETIC-STAGING FOR CANCER

Cancer patients with similar clinico-pathologic features (such as same stage of disease) could have markedly different treatment responses and disease outcomes (Lapointe and Li, 2004; van de Vijver et al., 2002). One of the reasons for this clinical heterogeneity is due to the genetic complexity of individual tumours, which changes in the expression of many genes that drive tumour growth, invasion and metastasis. In the past two decades, microarray technologies have transformed molecular genetics, allowing researchers to study the activity of thousands of genes simultaneously (McLachlan, Do, and Ambroise, 2004). There is an expectation that analysis of gene expressions will lead to better prognosis and the improvement of clinical outcomes by further our understanding of biological processes (Bullinger et al., 2004; Lapointe and Li, 2004) and more effective medication (appropriate doses or combinations of drugs) via pharmacogenomic profiling (Abbott, 2003; Halliday et al., 2004). Studies on molecular profiling (van't Veer et al., 2002; van de Vijver et al., 2002) indicate that breast cancer are heterogeneous and comprise diseases that are molecularly distinct and follow different clinical courses. Our study on four lung cancer datasets (Ben-Tovim Jones et al., 2005) also supports that the classification of patients based on gene expressions provides significant prognostic information on the disease outcome.

As described in Section 1, we propose to quantify this prognostic information from gene expressions as a genetic-staging for cancer. Among the large set of genes, it is anticipated that some genes may be redundant and have little information to separate the patients. Including such irrelevant genes can mislead classifications being made. A variable selection process can be used to determine particular subset of genes that are relevant. In this section, two mixture model-based clustering methods are considered to handle the dimensionality problem by identifying subset of genes that are relevant. The genetic stage for individual patient is determined on the basis of these selected genes.

2.1 Mixture model-based clustering methods

Finite mixture models have been widely applied in the fields of biomedical and health sciences as a device for clustering (McLachlan and Peel, 2000; Ng and McLachlan, 2004; Ng et al., 2004). In the present context of handling the dimensionality problem, a normal mixture model-based approach is adopted to cluster the genes based on the gene expression profiles. We let y_1, \dots, y_n denote n p -dimensional gene expression profile vectors. With this approach to clustering of gene expression data, it is assumed that each profile y_j is from a mixture of, say g , components of multivariate normal distributions in some unknown

mixing proportions π_1, \dots, π_g that sum to one. That is, y_j is taken to be a realization of the mixture probability density function (p.d.f.) defined by,

$$f(y, \Psi) = \sum_{i=1}^g \pi_i \phi(y; \mu_i, \Sigma_i), \quad (2.1)$$

where $\phi(y; \mu_i, \Sigma_i)$ denotes a multivariate normal p.d.f. with mean μ_i and covariance matrix Σ_i ($i = 1, \dots, g$). Here the vector Ψ of unknown parameters consists of the mixing proportions π_1, \dots, π_{g-1} , the elements of the component means μ_i , and the distinct elements of the component-covariance matrices Σ_i ($i = 1, \dots, g$). The maximum likelihood (ML) estimate of Ψ is obtained by application of the Expectation-Maximization (EM) algorithm of Dempster, Laird, and Rubin (1977); see also Ng, Krishnan, and McLachlan (2004). Within the mixture model framework for clustering, the value of g can be determined using the Bayesian information criterion (BIC) of Schwarz (1978). In addition, a probabilistic clustering of the genes into g components can be obtained based on the estimated posterior probabilities of component membership for the genes, $\tau_i(y_j; \hat{\Psi})$, where

$$\tau_i(y_j, \Psi) = \pi_i \phi(y_j; \mu_i, \Sigma_i) / f(y_j, \Psi) \quad (i = 1, \dots, g), \quad (2.2)$$

and $\hat{\Psi}$ denotes the ML estimate of Ψ . An outright assignment of the genes into g clusters is achieved by assigning each gene to the component to which it has the highest estimated posterior probability of belonging (McLachlan and Peel, 2000).

To handle the dimensionality problem, the first method adopts the first two steps of the software EMMIX-GENE (McLachlan, Bean, and Peel, 2002) that has been developed for a model-based approach to the clustering of microarray expression data. With EMMIX-GENE, the first step selects a subset of relevant genes by eliminating those genes which individually are of little use in clustering the tissue samples (patients) into groups, where the relevance of a gene is assessed on the basis of the likelihood ratio statistic for testing the number of groups in the mixture model (McLachlan, Bean, and Peel, 2002). This selection is undertaken without the known information on classification of tissue samples with respect to the disease outcome, and may therefore be considered as an unsupervised approach. The second step of EMMIX-GENE clusters the retained genes into $g = 20$ subgroups on the basis of Euclidean distance so that highly correlated genes are clustered into the same group. With reference to Equation 2.1, it means that a mixture in equal proportions $\pi_1 = \dots = \pi_g$ of normal distributions with covariance matrices restricted to being equal to a multiple of the identity matrix is being fitted (McLachlan, Bean, and Peel, 2002).

Each subgroup of genes is then represented by the sample mean expression profile, which is referred as a “meta-gene” for the subgroup (McLachlan, Do, and Ambroise, 2004).

In contrast to the first method, the second method adopts a supervised approach that makes use of the known classification of tissue samples on the disease outcome. The aim here is to identify “marker-genes” that are relevant to separate tissue samples into groups of different disease outcomes. The relevance of a gene is assessed on the basis of its level of differential expression between groups of known (pre-defined) disease outcomes (such as good- and poor-prognosis groups). The method adopts the EMMIX-WIRE approach of Ng et al. (2006) for the clustering of correlated gene-expression profiles. With this approach, the normal mixture model (Equation 2.1) is extended to capture correlations among genes via a linear mixed-effects model (McCulloch and Searle, 2001). Conditional on its membership of the i th component of the normal mixture, it is postulated that the distribution of y_j follows the model

$$y_j = X\beta_i + Ub_{ij} + Vc_i + \epsilon_{ij}, \quad (2.3)$$

where elements of β_i are fixed effects modelling the conditional mean of y_j in the i th cluster, the b_{ij} and c_i represent the unobservable random gene and tissue effects, and ϵ_{ij} is the measurement error vector (Ng et al., 2006). The random effects b_{ij} and c_i account for the variation due to the heterogeneity of genes and tissue samples, respectively. In Equation 2.3, X , U , and V are known design matrices for the corresponding fixed and random effects. With the linear mixed-effects model, the distributions of the random effects and the measurement error are taken to be multivariate normal with mean zero and unknown covariance matrix. We let $\Psi = (\varphi_1^T, \dots, \varphi_g^T, \pi_1, \dots, \pi_{g-1})^T$ be the vector of all the unknown parameters, where the superscript T denotes vector transpose and φ_i is the vector containing the unknown parameters β_i and the elements in the covariance matrices for the linear mixed-effects models ($i = 1, \dots, g$). The estimation of Ψ can be obtained by ML via the EM algorithm, proceeding conditionally on the tissue-specific random effects c_i as formulated in Ng et al. (2006). To identify marker-genes, a ranking of relevant genes can be obtained based on an estimated contrasts of fixed and the specific random effects b_{ij} for each component weighted by the estimated posterior probability of component membership (Ng et al., 2006).

2.2 Classification of patients into different genetic stages

The dimensionality problem for classifying patients into different genetic stages is handled by replacing the gene signatures of order over thousands by the top N meta-genes (the first method) or marker-genes (the second method). Patients are then classified into groups corresponding to different genetic stages, by fitting a normal mixture model (Equation 2.1) to the expression signatures of N dimensions corresponding to the N meta- or marker-genes. It is noted that the number of component (genetic stages) g is determined based on the BIC, as described in Section 2.1. That is, the method allows for a desirable property that the number of genetic stages can be different from the number of pre-defined groups corresponding to various disease outcomes.

The estimated normal mixture model can be used to allocate new or unclassified patients into one of the pre-determined groups of genetic stages. Under the mixture model approach to classification (McLachlan and Peel, 2000), the allocation is with respect to the components of the fitted mixture model. For the l th individual with gene signature vector y_l^* , we let $r(y_l^*) = i$ imply that y_l^* is assigned to the i th component ($i = 1, \dots, g$). The Bayes rule for the allocation of y_l^* is defined by

$$r(y_l^*) = i \quad \text{if} \quad \tau_i(y_l^*; \Psi) \geq \tau_h(y_l^*; \Psi) \quad (h = 1, \dots, g). \quad (2.4)$$

The Bayes rule can be estimated by the so-called plug-in rule, $r(y_l^*; \hat{\Psi})$, where $\hat{\Psi}$ denotes the ML estimate for the fitted normal mixture model (McLachlan and Peel, 2000). Assigning y_l^* to the group to which it has the highest posterior probability of belonging minimizes the expected misclassification rate (Fraley and Raftery, 2002).

3 PREDICTION OF DISEASE OUTCOME

Prediction of disease outcomes is then finalized by classifying patients using their genetic-staging as well as their corresponding clinical risk factors. In this study, logistic regression is adopted to relate the disease outcome with the genetic and clinical risk factors. Without the loss of generality, we assume that the response is a binary variable that indicates the membership of two pre-defined groups of disease outcomes corresponding to good- and poor-prognosis. Let x_l denote the observed vector of risk factors (genetic-staging and clinical factors), the logistic regression model is given by

$$\log(p_l/(1 - p_l)) = \alpha + \gamma x_l^T, \quad (3.1)$$

where p_l is the probability that the l th individual belongs to the group of good-prognosis, α is a constant term, and γ is a vector of parameters corresponding to the risk factors x_l . It is given within the generalized linear model (GLM) framework by specifying a binomial distribution function with the canonical logit link function. The unknown parameters (α and γ) are estimated by ML. From Equation 3.1, the log odds is taken to be linear. The odds ratios of good-prognosis for the risk factors are represented by the exponential of the corresponding ML estimates in $\hat{\gamma}$.

In this paper, we aim to compare the performance of prediction using clinical risk factors as well as the genetic-staging obtained from the gene expression profiles. The performance is based on the apparent error rates in each application to the observed data (that is, the proportion of the observations misallocated by the fitted logistic model). It is noted that the apparent error rates are considered here in a relative sense for the comparison. Caution should be exercised in interpreting these error rates in an absolute sense. This is because the apparent error rate is obtained by applying the fitted model to the same data from which it has been formed and hence provides an optimistic assessment of the actual error rates (McLachlan and Peel, 2000). In particular, the genetic-staging is obtained based on the meta- or marker-genes that are determined using the expression profiles from the cancer patients. Thus, the misclassification error rate is calculated without allowance for the selection bias (Ambroise and McLachlan, 2002). This bias may be corrected for the estimation of the error rate by using an “external” leave-one-out cross-validation, where the gene-selection procedure is performed at each stage of the cross-validation process on the remaining tissue samples; see McLachlan, Do, and Ambroise (2004).

4 EXAMPLE: BREAST CANCER DATA

The breast cancer gene expression data of van't Veer et al. (2002) is used in this illustration. The dataset consists of 5000 gene expression profiles and 6 binary variables of clinical indicators from 78 sporadic lymph-node-negative breast cancer patients; see the Supplementary Information of van't Veer et al. (2002) for the original coding of the 6 binary clinical variables. With these patients, 44 remained metastasis free after a period of more than 5 years (good prognosis) and 34 patients had developed distant metastases within 5 years (poor prognosis).

Based on the 5000 gene expression profiles, we obtained 20 meta-genes by the first clustering method described in Section 2.1. The top $N = 10$ meta-genes ranked in terms of the likelihood ratio statistic are used to classify patients into different genetic stages.

Table 1: Apparent error rates of classification (metastasis free or metastasis developed)

Method of classification based on	Misclassified (%)
clinical factors alone	20 (25.6%)
genetic-staging (meta-genes) alone	27 (34.6%)
genetic-staging (meta-genes) plus clinical factors	18 (23.1%)
genetic-staging (marker-genes) alone	22 (28.2%)
genetic-staging (marker-genes) plus clinical factors	16 (20.5%)

Based on the BIC for model selection (Schwarz, 1978), we identify there are two groups of patients corresponding to two different genetic stages (favourable and poor). The apparent error rates for predicting metastasis free based on clinical factors as well as genetic-staging are given in Table 1. For comparison, error rates based on clinical factors alone and genetic-staging alone are included. From Table 1, it can be seen that the error rate is smaller when classification is based on both genetic-staging and clinical factors (description of the results for genetic-staging obtained from marker-genes will be given in the next paragraph). In Table 2, the adjusted odds ratios and the corresponding 95% confidence intervals (CI's) of metastasis free are presented. It can be observed that the adjusted odds ratio of metastasis free is 2.8 (95% CI: 0.3 – 24.7) for patients with favourable genetic stage compared to those with poor genetic stage. This adjusted odds ratio, however, is not statistically significant at the 5% level. The only significant factor is a clinical one, Angioinvasion. From Table 2, patients without angioinvasion have a higher chance of being metastasis free compared to those with angioinvasion (adjusted odds ratio= 4.4, 95% CI: 1.3 – 15.2).

The second clustering method (Section 2.1) is then applied to identify marker-genes based on the 5000 gene expression profiles. The top $N = 10$ marker-genes ranked in terms of the differential expression between the two prognosis groups are used to classify patients into groups of genetic stages. Based on the BIC for model selection, we identify there are three groups of patients corresponding to three genetic stages (favourable, mild, and poor). The apparent error rates for predicting metastasis free are presented in Table 1 as well. It can be observed that genetic-staging obtained from marker-genes leads to a smaller error rate, compared to that obtained from meta-genes. This is anticipated as the former makes use of the known classification of tissue samples on the disease outcome. The adjusted odds ratios (and the 95% CI's) of metastasis free are presented in Table 3. The

Table 2: Adjusted odds ratio of metastasis free (genetic stage from meta-genes)

Variable	Odds ratio (95% CI)
Tumour grade (1 or 2 vs 3)	2.9 (0.8 – 11.0)
Oestrogen receptor status (≤ 10 vs > 10)	1.1 (0.1 – 9.9)
Progesteron receptor status (≤ 10 vs > 10)	1.2 (0.2 – 5.7)
Tumour size (≤ 20 vs > 20 mm)	2.6 (0.8 – 8.1)
Patient age (≤ 40 vs > 40)	0.3 (0.09 – 1.1)
Angioinvasion (no vs yes)	4.4* (1.3 – 15.2)
Genetic stage from meta-genes (favourable vs poor)	2.8 (0.3 – 24.7)

* significant result at the 5% level.

genetic-staging is statistically significant when it is obtained from marker-genes. From Table 3, patients having a favourable genetic stage are more likely being metastasis free compared to those having a poor genetic stage (adjusted odds ratio= 23.0, 95% CI: 3.0 – 178). Similarly, patients with a mild genetic stage also have a higher chance of being metastasis free compared to those with a poor genetic stage (adjusted odds ratio= 6.0, 95% CI: 1.3 – 27.7). The clinical factor Angioinvasion is now only marginally significant.

5 DISCUSSION

We have presented two mixture model-based clustering methods to quantify the prognostic information from gene expressions as a genetic-staging for cancer. This is proceeded by classifying patients into groups based on the top N meta-genes (the first clustering method) or marker-genes (the second clustering method). The groups so obtained correspond to patient groups of different genetic stages. The impacts of genetic-staging and clinical risk factors on the prediction of disease outcomes are then assessed using logistic regression model. In Section 4, the proposed method is illustrated using a real example of breast cancer data. From Tables 1 and 3, it can be seen that genetic-staging, obtained by applying sophisticated model-based clustering method for the identification of marker-genes that are relevant to disease outcome, can provide significant additional prognostic information. A reliable and precise prediction of individual disease outcomes is essential to identify high-risk individuals for more aggressive surgical and adjuvant treatment while

Table 3: Adjusted odds ratio of metastasis free (genetic stage from marker-genes)

Variable	Odds ratio (95% CI)
Tumour grade (1 or 2 vs 3)	2.8 (0.7 – 12.0)
Oestrogen receptor status (≤ 10 vs > 10)	0.3 (0.05 – 1.8)
Progesteron receptor status (≤ 10 vs > 10)	1.8 (0.3 – 9.7)
Tumour size (≤ 20 vs > 20 mm)	2.8 (0.8 – 9.7)
Patient age (≤ 40 vs > 40)	0.4 (0.09 – 1.6)
Angioinvasion (no vs yes)	4.2* (1.0 – 17.1)
Genetic stage from marker-genes (favourable vs poor)	23.0* (3.0 – 178)
(mild vs poor)	6.0* (1.3 – 27.7)

* significant result at the 5% level.

avoiding unnecessary treatment with toxic side-effects in good-risk patients (Shaha, 2004). The study is therefore relevant for its attempt to bridge the gap between public health and genetics towards the goal of improved disease outcomes and more targeted treatment and health care for cancer patients.

In the illustration, we consider the simultaneous use of genetic-staging and 6 clinical factors for the prediction of disease outcomes. The proposed method is indeed applicable to handle other types of patient's characteristics that may have significant impacts on predicting disease outcomes. This includes patient's demographic, socio-economics, and behavioural factors (such as patient age, alcohol consumption, and smoking behaviour). Recent study has shown that high alcohol consumption and smoking are non-genetic risk factors associated with colorectal cancer in men (Otani et al., 2003); see also Ferrai et al. (2007). In addition, epidemiologic studies on ovarian cancer (Green et al., 2001; Zhang et al., 2004) have indicated that women with smoking exposure have significant higher risk of developing mucinous ovarian cancer. Thus, the inclusion of these non-genetic risk factors into the model (Equation 3.1) may help to improve further the prediction of disease outcomes and identify significant "avoidable" risk factors (such as smoking exposure), which may be used to determine potential public health intervention programs for the disease (Brand, Brand, and den Baumen, 2008; Green et al., 2001).

In this paper, the number of top ranked meta-genes (or marker-genes), N , is chosen arbi-

trarily. In practice, an external cross-validation procedure may be used to select an optimal value for N that minimizes the misclassification error rate; see Ambroise and McLachlan (2002) and McLachlan, Do, and Ambroise (2004) for the details. Alternatively, we may choose a reasonably large value for N (say, $N = 100$) and then adopt mixtures of factor analyzers to classify patients into different genetic stages on the basis of the N selected meta- (or marker-) genes. The use of mixtures of factor analyzers handles the dimensionality problem by imposing the assumption that the correlations between the genes can be expressed in a lower space by the dependence of the tissue samples on q ($q \ll N$) unobservable factors (McLachlan, Do, and Ambroise, 2004).

5.1 Acknowledgements

This work is supported by a research grant from the Griffith University, Australia.

REFERENCES

1. Abbott, A. (2003). With your genes? Take one of these, three times a day. *Nature*, **425**, 760-762.
2. Ambroise, C. and McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene expression data. *Proc. Natl. Acad. Sci. USA*, **99**, 6562-6566.
3. Ben-Tovim Jones, L., Ng, S.K., Ambroise, C., Monico, K., Khan, N. and McLachlan, G.J. (2005). Use of microarray data via model-based classification in the study and prediction of survival from lung cancer. In *Methods of Microarray Data Analysis IV*, J.S. Shoemaker and S.M. Lin (Eds.). Springer, New York, pp. 163-173.
4. Brand, A., Brand, H. and den Baumen, T.S. (2008). The impact of genetics and genomics on public health. *Euro. J. Human Genetics*, **16**, 5-13.
5. Bullinger, L., Dohner, K., Bair, E., et al. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N. Engl. J. Med.*, **350**, 1605-1616.
6. Caldas, C. and Aparicio, S.A.J. (2002). Cancer: The molecular outlook. *Nature*, **415**, 484-485.

7. Carlsten, C. and Burke, W. (2006). Potential for genetics to promote public health – Genetics research on smoking suggests caution about expectations. *J. Am. Med. Assoc.*, **296**, 2480-2482.
8. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1-38.
9. Ferrari, P., Jenab, M., Norat, T., et al. (2007). Lifetime and baseline alcohol intake and risk of colon and rectal cancers in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Int. J. Cancer*, **121**, 2065-2072.
10. Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611-631.
11. Green, A., Purdie, D., Bain, C. Siskind, V. and Webb, P.M. (2001). Cigarette smoking and risk of epithelial ovarian cancer (Australia). *Can. Causes Control*, **12**, 713-719.
12. Halliday, J.L., Collins, V.R., Aitken, M.A., Richards, M.P.M. and Olsson C.A. (2004). Genetics and public health – evolution or revolution? *J. Epidemiol. Community Health*, **58**, 894-899.
13. Kim, S., Tadesse, M.G. and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93**, 877-893.
14. Kudo, M. and Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, **33**, 25-41.
15. Lapointe, J. and Li, C. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA*, **101**, 811-816.
16. McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.
17. McLachlan, G.J., Bean, R.W. and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413-422.
18. McLachlan, G.J., Do, K.-A. and Ambrose, C. (2004). *Analyzing Microarray Gene Expression Data*. Wiley, New Jersey.
19. McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

20. Ng, S.K., Krishnan, T. and McLachlan, G.J. (2004). The EM algorithm. In *Handbook of Computational Statistics Vol. 1*, J. Gentle, W. Hardle and Y. Mori (Eds.). Springer-Verlag, New York, pp. 137-168.
21. Ng, S.K. and McLachlan, G.J. (2004). Speeding up the EM algorithm for mixture model-based segmentation of magnetic resonance images. *Pattern Recognition*, **37**, 1573-1589.
22. Ng, S.K., McLachlan, G.J., Wang, K., Ben-Tovim, L. and Ng, S.W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, **22**, 1745-1752.
23. Ng, S.K., McLachlan, G.J., Yau, K.K.W. and Lee, A.H. (2004). Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. *Stat. Med.*, **23**, 2729-2744.
24. Nguyen, D.V. and Rocke, D.M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.
25. Otani, T., Iwasaki, M., Yamamoto, S., et al. (2003). Alcohol consumption, smoking, and subsequent risk of colorectal cancer in middle-aged and elderly Japanese men and women: Japan public health center-based prospective study. *Cancer Epidemiol. Biomark. Prev.*, **12**, 1492-1500.
26. Rhodes, K.V. and Pollock, D.A. (2006). The future of emergency medicine public health research. *Emerg. Med. Clin. N. Am.*, **24**, 1053-1073.
27. Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**, 461-464.
28. Shaha, A.R. (2004). Implications of prognostic factors and risk groups in the management of differentiated thyroid cancer. *Laryngoscope*, **114**, 393-402.
29. van't Veer, L.J., Dai, H., van de Vijver, M.J., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530-536.
30. van de Vijver, M.J., He, Y.D., van't Veer, L.J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999-2009.

31. West, M., Blanchette, C., Dressman, H., et al. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA*, **98**, 11462-11467.
32. Xiong, M., Li, W., Zhao, J., Jin, L. and Boerwinkle, E. (2001). Feature (gene) selection in gene-expression-based tumor classification. *Mol. Genet. Metab.*, **73**, 239-247.
33. Zhang, Y., Coogan, P.F., Palmer, J.R., Strom, B.L. and Rosenberg, L. (2004). Cigarette smoking and increased risk of mucinous epithelial ovarian cancer. *Am. J. Epidemiol.*, **159**, 133-139.