Independent Component Analysis of Microarray Data in the Study of Endometrial Cancer

(Brief Title: Independent Component Analysis for Gene Arrays)

Samir A. Saidi¹, Cathrine M. Holland⁴, David P. Kreil²³, David J.C. MacKay², D. Stephen Charnock-Jones¹⁴, Cristin G. Print⁴, Stephen K. Smith¹⁴

¹ Department of Obstetrics and Gynaecology, University of Cambridge, Cambridge, CB2 2SW, UK
² Inference Group, Department of Physics, Cavendish Laboratory, Cambridge, CB3 0HE, UK
³ Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
⁴ Department of Pathology, University of Cambridge, Cambridge, CB1 1QP, UK

Corresponding author:
Samir A Saidi
University Department of Obstetrics and Gynaecology
Box 223, The Rosie Hospital
Robinson Way,
Cambridge,
CB2 2SW
Phone: +44 1223 336874
Fax: +44 7010 707447
Email: samsaidi@obgyn.cam.ac.uk
ABSTRACT

Gene microarray technology is highly effective in screening for differential gene expression, and has hence become a popular tool in the molecular investigation of cancer. When applied to tumours, molecular characteristics may be correlated with clinical features such as response to chemotherapy. Exploitation of the huge amount of data generated by microarrays is difficult, however, and constitutes a major challenge in the advancement of this methodology. Independent Component Analysis (ICA), a modern statistical method, allows us to better understand data in such complex and noisy measurement environments. The technique has the potential to significantly increase the quality of the resulting data, and improve the biological validity of subsequent analysis. We performed microarray experiments on 31 postmenopausal endometrial biopsies, comprising 11 benign and 20 malignant samples. We compared ICA to the established methods of Principal Component Analysis (PCA), Cyber-T, and SAM. We show that ICA generated patterns that clearly characterised the malignant samples studied, in contrast to PCA. Moreover, ICA improved the biological validity of the genes identified as differentially expressed in endometrial carcinoma, compared to those found by Cyber-T and SAM. In particular, several genes involved in lipid metabolism that are differentially expressed in endometrial carcinoma were only found using this method. This report highlights the potential of ICA in the analysis of microarray data.
In the last 3 years there has been an explosion in the number of studies involving the use of gene microarrays. Microarrays give researchers the ability to investigate hundreds to thousands of genes in parallel. Microarray technology allows for two main types of descriptive analysis: firstly, the identification of genes which may be responsible for a clinicopathological feature or phenotype (using the array as a screening tool) and, secondly, the genomic classification of tissue. The ultimate goal is to improve clinical outcome by adapting therapy based on the molecular characteristics of a tumour. Initially, the identification of individual genes by microarray screening provides promising candidates for targeted research. Furthermore, identifying overall patterns of gene changes can reveal candidates that can be missed in a gene-by-gene approach.

Progress using this technique, however, has been hampered by issues of reproducibility other than the expected biological variation in gene expression. Due to the inherent noise (stochastic error) in most current array systems, single control-vs-experiment designs are unreliable. Overcoming this requires a combination of solid experimental design, replicate samples, and more sophisticated statistical tools than have traditionally been applied. Preferably, microarray experiments might employ ANOVA-based methods allowing for multiple confounders, but this requires costly experimental replication.

Attention has mostly therefore been focussing on the post-analytical tools used to process the microarray data. Sophisticated ‘single-gene’ methods such as a Bayesian t-test, ‘Cyber-T’ and SAM have been developed, which improve the reliability of significance analysis for individual genes.

Multi-gene approaches, which aim to detect patterns of gene co-variation, are more promising from a biological perspective. Clustering-based visual tools, such as hierarchical clustering, divisive clustering and self-organising maps, have been popular methods in this field, grouping together genes with similar patterns of expression. Such methods, however, typically assume that each gene belongs to just one cluster. This assumption fails where
genes belong to two or more independent expression patterns, as would be the case if a gene were influenced by multiple transcription factors, or was involved in more than one pathway.

Principal component analysis (PCA)\textsuperscript{13} takes a different approach trying to identify components which explain the variance in the data. Genes of interest may then be inferred from genes constituting the components found to correlate with the subject of investigation. The constraint of mutual orthogonality, of components implied by classical PCA, however, may not be appropriate for the biological systems studied. A biologically more plausible model is assumed by Independent Component Analysis (ICA), in which the components are by design statistically independent, a much weaker constraint. There is as yet little published data on the use of ICA for analysis of microarray data, particularly relating to tumour data\textsuperscript{14,15}.

We have adopted an Ensemble Learning implementation of ICA\textsuperscript{16}, which we use to identify latent variables underlying the data. In particular, this technique also allows us to estimate the inherent measurement noise. The model for tissues $t$, gene signals $g$, and components $m$ is of the form

$$s_{tg} = \sum a_{tm} b_{mg} + n_{tg}$$

or, in matrix notation,

$$S = AB + N,$$

where $B$ contains the components (representing gene signatures) and $A$ the amounts (or weights) of those components present in each tissue. The matrix $N$ represents the noise in the data. All three are estimated from the observed signal matrix $S$. 
This model enables us to identify and subsequently remove, or filter, measurement noise from the data set and therefore improve the biological relevance of the gene expression patterns detected. We exploit the fact that ICA can separate the patterns in which we are interested from independent other effects, like random sample variation or biological patterns unrelated to the subject of investigation. One such pattern is seen in figure 1 as the first component. This component seems to have no biological relevance to cancer, as its value is similar across all samples. We can reconstruct the data with both noise and this extraneous component removed: the ‘common’ component is subtracted from the amounts matrix $A$ to give the new matrix $A'$, where $a'_{tm} = a_{tm}$ for $m \neq 1$, and $a'_{t1} = 0$. After subtraction of the noise matrix $N$, the newly reconstructed data $S_R$ are then given by

$$S_R = A'B.$$  \hspace{1cm} (3)

To evaluate ICA applied to tumour gene expression data, we performed microarray analysis on a set of 31 endometrial tissue specimens (11 benign and 20 malignant samples comprising 17 endometrioid carcinomas and 3 serous papillary carcinomas). We used nylon membrane-based microarrays designed in-house with 1056 duplicated probes giving a total of 2112 spots (known genes, ESTs, and some calibration spikes). The choice of nylon-based over glass arrays was made due to their superior reproducibility in this laboratory \cite{17}.

**ICA improved clustering patterns of tumour microarray data**

Unsupervised hierarchical clustering was applied to the original normalised data, the data as reconstructed by ICA, and the data in independent component space (the amounts matrix $A$). Figure 2 demonstrates the filtering capacity of the ICA model. For the original data, there was some clustering following the expected pattern, in that most of the malignant samples have clustered together, but the highest hierarchical split did not separate the groups as would have been expected (Figure 2a). For the data reconstructed by ICA through
removal of the noise matrix $N$ and subtraction of the ‘common’ component, results overall demonstrated clear separation of the benign and malignant groups (Figure 2b), which is a considerable improvement over the original.

Figure 1a shows the results of the same clustering algorithm applied to the amounts matrix $A$ derived by ICA. As can be seen from the corresponding hierarchical cluster dendrogram, the ICA algorithm extracted the characteristic expression patterns corresponding to the histological classification in this reduced data set. The clustering now fully discriminates between the benign and malignant endometrial tissues. In effect, ICA has filtered-out the measurement noise whilst at the same time maintaining the relevant structure of the data, in a fraction of the size of the original matrix. The clustering pattern shown was achieved using the amounts matrix $A$ for all 31 derived independent components. Clustering based on the first 15 or more independent components, furthermore, produced an identical pattern, suggesting that these components alone capture sufficient biologically relevant information. This is contrasted by clustering of the original data, where the biologically relevant patterns were obscured by noise. Classical PCA suffered similarly when applied to the original, noisy data. In clustering after PCA, the pattern closest to the histological classification was produced when restricting the analysis to the first 3 or 4 principle components. Despite this, PCA failed to produce a pattern matching the histological classification (Figure 1b).

**ICA identified novel patterns of co-regulated genes clinically associated with endometrial carcinoma**

Once a component derived by ICA has been identified as correlated to the subject of interest, this also implicates those individual genes that give the strongest contribution to that component (fig.1). To establish whether ICA improved the biological relevance of candidate target genes, we compared the characteristics of genes identified by ICA to those found using the more established single-gene methods Cyber-T and SAM.
Figure 4 shows the chromosome distribution of the highly significant gene spots that had been selected for the comparison of ICA, Cyber-T and SAM. In contrast to Cyber-T and SAM, ICA identified a significantly higher number of genes on chromosome 17 than would be expected from the overall chromosomal distribution of genes on the array. More specifically, of the genes identified by ICA, 10 hits were from genes mapping to cytoband 17q21, a locus which has previously been shown to exhibit Loss of Heterozygosity in endometrial cancer.18,19 Cyber-T and SAM, however, identified only 6 and 2 gene hits, respectively, at the 17q21 locus.

Genes were grouped by function according to Gene Ontology annotation. For each functional group, its prevalence amongst the genes identified by ICA, Cyber-T, or SAM was then compared to its prevalence amongst all arrayed genes, using a chi-squared test. ICA identified significantly increased numbers of genes relating to two important functional groups. All three methods had identified increased numbers of genes involved with insulin-like growth-factor receptor binding (GO codes 5067, 5159, 5010, 5520; p<0.05). Genes derived using ICA additionally showed a significantly increased prevalence of functional groups related to lipid and fatty acid metabolism (GO codes 6629, 6631; 3.6% vs 0.5%, p<0.0001). The corresponding analysis of the genes derived from both Cyber T and SAM, however, failed to reveal this salient feature (0.9% and 0.8%, respectively, vs 0.5%, p>0.5).

**Biological validity and robustness**

A further corroboration of the validity of the ICA results was facilitated by the design of the microarray, as gene probes were spotted onto the array in pairs. These gene pairs can be used as controls because the knowledge about the pairing of the gene spots has not been
used in the data model of the original analysis. When gene list selection was based on only the top-ranked 2% of selected genes, Cyber-T and ICA produced the same high proportion of paired genes, confirming both methods as reliable. As soon as more of the top ranking genes were examined, however, ICA produced significantly more paired gene spots than either of the other methods (for the top-ranked 5% of selected genes: ICA 75%, vs Cyber-T 53% vs SAM 14%, p<0.001, chi-squared test; for the top 2%: ICA/Cyber-T 71%, SAM 38%)

Moreover, ICA improved the robustness of subsequent analyses. Leave-one-out analysis was conducted by in turn dropping data from one tissue sample from the entire data set before clustering. Initially, either the raw normalised data, or the corresponding data after filtering using ICA were examined. The mean number of misclassifications by clustering was significantly lower using the ICA-derived data instead of the raw data (2.74 vs 4.97, p<0.01, Mann-Whitney test).

In addition, only the data for the lists of genes selected by ICA, Cyber-T, and SAM were clustered. In leave-one-out analysis, the data for the gene list derived using ICA again showed the lowest number of misclassifications by clustering (0.6 for ICA vs 1.0 for Cyber-T and 3.1 for SAM, p<0.001). Moreover, data for the ICA gene list much more often yielded clustering patterns with zero misclassifications (21/31=68% vs 3/31=10% for Cyber-T and 2/31=6% for SAM, p<0.0001, chi-squared test).

**Increased biological relevance in microarray analyses using ICA**

In this study, Independent Component Analysis (ICA) identified co-regulated genes associated with endometrial cancer. Traditionally, the pathogenesis of this tumour is
considered to be related to the effects of unopposed estrogenic stimulation of the uterus\textsuperscript{22}. There are, however, strong associations with obesity and hyperinsulinaemia\textsuperscript{22}, and insulin-like growth factor has been shown to be mitogenic in endometrial cancer-derived cell lines in vitro\textsuperscript{23}.  

There is increasing evidence that lipid metabolism pathways are highly influential in the development and maintenance of some cancers\textsuperscript{24}. Considering the established association with obesity, it is highly plausible that genes involved in lipid metabolism are implicated in endometrial carcinogenesis. This is the first study reporting molecular evidence for such a link.  

Of the methods compared, only ICA successfully identified both the already established connection of endometrial cancer with insulin-like growth factor receptor binding, and the novel link to lipid metabolism.  

ICA removed noise and independent patterns unrelated to the subject of investigation. Consequently, results of analyses employing ICA in general showed increased biological relevance and robustness. This was examined using clustering analysis, which showed marked improvements of the clusters matching the histological classification of the samples, for either the data filtered using ICA or the independent component amounts. The reliable classification of tumours based on their transcript pattern is of particular clinical interest, yet validation of predictive classification will require larger datasets. The improvements seen in this study as the result of removing confounding effects, however, are likely to benefit any subsequent data analysis.  

In summary, the ability of ICA to extract biologically relevant gene expression information from microarray data, despite the presence of significant noise, highlights the potential of this method for microarray analysis.
Acknowledgments

We thank Dr Andrew Sharkey and Amanda Evans for their assistance in design and construction of the arrays used in these studies, and Addenbrooke’s hospital, Cambridge, for their collaboration. This work was carried out in the MRC Reproductive Angiogenesis Cooperative Group and was supported in part by program grant G9623012. C. Holland acknowledges support from an MRC Research Training Fellowship (G84/5733) and the Raymond and Beverly Sackler Studentship Award 2002. D. Kreil acknowledges support by an MRC Research Training Fellowship (G81/555).

Data availability

Further data supporting this manuscript and technical details are available in an online supplement (http://www.obgyn.cam.ac.uk/genearray/supplementary)
FIGURE LEGENDS

Figure 1
(a) ICA output with superimposed hierarchical cluster dendrogram. The columns represent endometrial samples (B1-B11 benign, M1-M20 malignant). The independent components are shown as rows and have been ordered by data power which is also a good indicator of component robustness. The second component shows a ‘Cancer’ gene signature (see below and figure 3). The size of each square corresponds to the amount $a_{lm}$ of component $m$ in tissue $t$. Red and yellow represent positive and negative values respectively. Hierarchical clustering was applied to the columns of the data matrix and showed a distinct separation into benign groups (green lines) and malignant groups (blue lines). The same clustering pattern was obtained when applied to the first 15 or more components. (b) PCA applied to the same data set. The clustering patterns shown are based on (i) $k=15$ and (ii) $k=4$ components. For PCA, the clustering pattern closest to the histological classification was achieved with $k=4$ components.

For the ICA-derived data, the amounts $a_{lm}$ were, for each component $m$, compared by histological classification of the tissues using a Mann-Whitney U-test. Only the second component was relevant to the classification into benign and malignant tissues, and was therefore selected for further analysis. The significance of this ‘cancer’ component was validated by bootstrap resampling. The data were permuted randomly for each array and the reordered data sets were subjected to ICA. The amounts for each component were then compared by histological classification of the tissues using a Mann-Whitney U-test, recording the most significant p-value achieved. This process was independently repeated 1000 times. The distribution of the observed extreme $\log_{10} p$-values was compact and unimodal, with median=-0.15, median absolute deviation=0.22, and minimum=-4.1. The observed
distribution confirmed the original ICA result to be robust and expected <0.1% by chance. Conservative correction for multiple testing was applied to all quoted p-values.

**Figure 2:** (a) Clustering of raw (normalised) data shows poor correlation with the histological classification. (b) The same data reconstructed following ICA. An improvement in clustering pattern was achieved by removal of the noise from the data set. While removal of components unrelated to the subject of interest will in general improve subsequent analysis, subtraction of the 'common' component in this particular case only improved the clarity of the expression patterns, not the clustering dendrogram. 2-way hierarchical clustering has been applied, by experiments and genes, with subsequent ordering of the data to the clusters. B1-11: benign samples, M1-20: malignant. Red/Green blocks represent signal increase/decrease respectively from the control log mean - signal changes greater than 2 standard deviations show the brightest coloration.

Hierarchical clustering of the raw data \( S \), the ICA reconstructed data \( S_R \), and the reduced data from ICA (\( A' \)) and also from PCA was performed using Ward’s method with Euclidean distances\(^{26} \). Other popular clustering methods were also tested, and gave similar results.

**Figure 3:** Gene signature of the 'cancer' component (row 2 from figure 1a). Genes with loadings exceeding the chosen percentile lines were considered significant. Positive (red) and negative (yellow) loadings correspond to up- and down-regulation of expression, respectively. Ensemble learning provides results with error bars, which were consistently < 0.07 for this component, and have not been plotted for clarity.

The ICA algorithm computed the loading \( b_{mg} \) for each gene signal \( g \) in component \( m \) (\( g = 1..2112, m = 1..31 \)). The loading \( b_{mg} \) represents the relative contribution of the gene signal \( g \)
to the component \( m \). The loadings \( b_{mg} \) of component 2 (figure 1) were plotted with threshold lines at the 95\(^{th}\) percentiles for positive and negative values.

**Figure 4**: Distribution of hits (gene spots identified as significant) after analysis using ICA (yellow), Cyber-T (blue) and SAM (pale blue bars). The 'hit density' is the ratio of the number of hits to the relative chromosome length. The distribution of gene spots (by chromosome) on the array is represented by black bars for comparison. The error bars represent the 95\% upper limit of expected hits for each chromosome (based on the binomial probabilities). The number of hits to chromosome 17 returned by ICA was significantly higher than that expected from the overall chromosomal distribution of genes on the array (14/89 vs 116/1740 spots, \( p<0.05 \), chi-squared test). In contrast, the corresponding analysis for the genes selected by Cyber-T and SAM failed to reach significance (Cyber-T: 8/82, SAM: 8/88; vs 116/1740, \( p>0.2 \)).

Corresponding thresholds were chosen for Cyber-T and SAM, respectively, in order to allow a direct comparison of the methods at the gene level. In Cyber-T, these genes all had Bayesian \( p \)-values < 0.01. The web interface for Cyber-T\(^{11} \) was used for analysis with a sliding-window size of 61 (approximately 3\% of the spots). Default parameters were chosen in the analysis using SAM for two-class data. We combined the patterns of genes derived from ICA with annotation from public genomic database sources (Ensembl, Unigene, Locuslink, GO)\(^{20,27,28} \). Somatic cytoband locations were found for 870 genes (1740 spots). A total of 6179 functional groups could be assigned by ontology to 895 genes (1790 spots).
References


