BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 8 2007, pages 957–965
doi:10.1093/bioinformatics/btm033

Gene expression

Disease-specific genomic analysis: identifying the signature of pathologic biology

Monica Nicolau[1,2], Robert Tibshirani[3,4], Anne-Lise Børresen-Dale[5,6] and Stefanie S. Jeffrey[1,*]

[1]Department of Surgery, Stanford University School of Medicine, [2]Department of Mathematics, [3]Department of Health, Research & Policy, [4]Department of Statistics, Stanford University, [5]Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Center and [6]Medical Faculty, University of Oslo, Oslo, Norway

Received on May 9, 2006; revised on December 22, 2006; accepted on January 28, 2007

Advance Access publication February 3, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Genomic high-throughput technology generates massive data, providing opportunities to understand countless facets of the functioning genome. It also raises profound issues in identifying data relevant to the biology being studied.

Results: We introduce a method for the analysis of pathologic biology that unravels the disease characteristics of high dimensional data. The method, *disease-specific genomic analysis* (*DSGA*), is intended to precede standard techniques like clustering or class prediction, and enhance their performance and ability to detect disease. *DSGA* measures the extent to which the disease deviates from a continuous range of normal phenotypes, and isolates the aberrant component of data. In several microarray cancer datasets, we show that *DSGA* outperforms standard methods. We then use *DSGA* to highlight a novel subdivision of an important class of genes in breast cancer, the *estrogen receptor* (ER) cluster. We also identify new markers distinguishing *ductal* and *lobular* breast cancers. Although our examples focus on microarrays, *DSGA* generalizes to any high dimensional genomic/proteomic data.

Contact: ssj@standford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The genomic era has brought about profound changes in the study of genetic mechanisms with the infusion of mathematical tools to aid both traditional and novel biological techniques. High dimensional data like microarray expression, *SNP*, array *CGH* and proteomic data have been used to study a wide range of problems aimed at achieving a deeper and more global understanding of diseases. Identification of expression relevant to the biological problem being studied can however be a difficult task. Tests for statistical significance must always make tacit assumptions about the underlying biology, and different tests will highlight distinct aspects of this biology. In studies of diseases, statistical analysis is often employed to identify the most important variables (genes). This most commonly includes genes that vary a lot among distinct tumors (Alon *et al.*, 1999; Dudoit Fridlyand *et al.*, 2002; Eisen *et al.*, 1998; Golub *et al.*, 1999; Hastie *et al.*, 2000; Weinstein *et al.*, 1997), genes whose expression is stable among different samples from the same patient (Sørlie *et al.*, 2003; Weigelt *et al.*, 2005), genes whose expression levels show a strong association with various clinico-pathologic characteristics (Bair *et al.*, 2006; Dudoit Yang *et al.*, 2002; Tusher *et al.*, 2001; Vijver *et al.*, 2002; Weigelt *et al.*, 2005), and various methods that identify genes whose expression most significantly distinguish diseased and normal tissues: (Alon *et al.*, 1999; Boer *et al.*, 2001; Chen *et al.*, 2002; Ghosh *et al.*, 2004; Munagala *et al.*, 2004; Stephanopoulos *et al.*, 2002).

In this article we introduce a novel method of data analysis: *disease-specific genomic analysis* (*DSGA*) that employs comparison to normal expression to extract data most closely associated with the disease. Specifically, *DSGA* defines a supervised step that mathematically transforms and simplifies expression data to highlight the pathologic component of expression. While retaining expression information about *every* gene, *DSGA* isolates and separates a disease-like and a normal-like portion of this expression. Other, standard analytic methods—clustering, class prediction, feature selections—are meant to be applied *after* the data has been transformed by *DSGA*. This method defines the mathematical model for *normal* expression to be a linear subspace derived from normal tissue expression data; it defines *disease-specific* expression to be the deviation of expression in diseased tissue from this subspace, where *deviation* indicates residual from a linear model. Specifically, we define a subspace $\mathcal{N}$ that approximates normal tissue data (Section 2.2), and then decompose the original expression data T from each individual diseased tissue into two components: the *normal component* $Nc.T$ is the least squares fit of T to a linear model in $\mathcal{N}$, and the *disease component* $Dc.T$ is the vector of residuals from the fit

*To whom correspondence should be addressed.

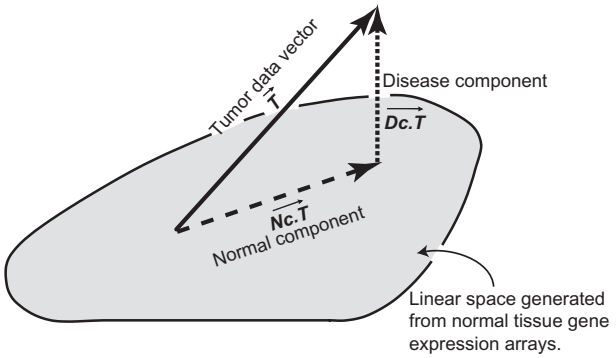

Fig. 1. Geometric representation of the decomposition of *tumor* data vector into *normal component* and *disease component* vectors. The linear space $\mathcal{N}$ is obtained from the normal tissue data using the method defined in Section 2.

to this linear model. The two vectors $Nc.T$ and $Dc.T$ are perpendicular, and satisfy:

$$T = Nc.T + Dc.T \quad (1)$$

This construction allows each diseased tissue expression vector to find its own unique normal component (fit to a linear model) to the normal state. Figure 1 shows a geometric representation of the different components. Standard data analysis methods are subsequently applied to the disease components $Dc.T$ of the data.

The method, detailed in Section 2.2, involves essentially computing the residual deviation of diseased tissue data from a linear model in the normal tissue data. A modification of principal component analysis is used to obtain a good approximation of the model for normal expression. This method is tested in Section 2.3 using data simulations.

In Section 3, we apply *DSGA* to real microarray expression data to investigate the benefits it provides. Specifically, we show:

(1) *DSGA* outperforms standard analysis methods by accurately recognizing clinical, a priori established biology with improved error rates.

(2) *DSGA* tends to highlight aspects of biology that are distinct from those identified by traditional methods. Hence the *DSGA* decomposition has the potential to identify novel biology, rather than uncover, albeit with improved accuracy, essentially known biological identities.

We note that the second statement does *not* make the claim that the biology identified using *DSGA* is more accurate, or more revealing than the biology highlighted by traditional methods. It merely states that this biology is different. The first statement however, claims that *DSGA* decomposed data is better at correctly identifying a priori known biology than traditional data.

In Section 3.1, we show that *DSGA* decomposition of diseased tissue data performs better than the original (log ratio) data for class prediction by *prediction analysis for microarrays* (*PAM*) (Tibshirani *et al.*, 2002) by testing on several cancer datasets. Specifically, performing *DSGA* transformation on log ratio data places tumors in classes defined by clinico-pathological parameters with better error rates. Indeed, this suggests that this transformation highlights the characteristics of data that are relevant to the biology of disease, and that other traditional analysis methods should be applied to *DSGA*-transformed data rather than to the original data.

In Section 3.2, we investigate the second question: to what extent is *DSGA* likely to uncover new aspects of biology. We focus on breast cancer, where we highlight two separate instances where differences between *DSGA* and other methods are clearly discernable. The first difference concerns the predictor genes identified in the *PAM* analysis to distinguish *ductal* and *classical lobular* breast cancer tumors. Thus while in Section 3.1 we show that error rates for *PAM* are improved when using *DSGA*-decomposed data, in Section 3.2 we show that this same *PAM* analysis has identified a different collection of predictor variables (genes) in constructing the tumor class shrunken centroids. Thus not only is the error rate improved, but the predictor genes are different, thereby potentially uncovering novel biology. Second, we use the disease components of *DSGA*-transformed data to highlight novel gene associations for breast cancer; specifically we discover a decomposition of the *estrogen receptor (ER)* cluster into three subclusters of biologically coherent gene groups that are associated with distinct tumor types. Given the long recognized biological importance of *ER* status in the development and progression of breast cancer (Creighton *et al.*, 2006; Foekens *et al.*, 2006; Gruvberger *et al.*, 2001; Innes *et al.*, 2006; Laganiere *et al.*, 2005; Oh *et al.*, 2006; Paik *et al.*, 2004; Perou *et al.*, 2000; Sørlie *et al.*, 2001; Sørlie *et al.*, 2003; Usary *et al.*, 2004; Wang *et al.*, 2005; Yang *et al.*, 2006) this finding highlights the potential value of *DSGA* in further unraveling the underlying biology in disease. In Section 4, we discuss some characteristics of data decomposition by *DSGA*.

2 DISEASE-SPECIFIC GENOMIC ANALYSIS

Our method is based on decomposing expression in diseased tissue as the sum of a part that best mimics normal tissue expression, and an error or deviation from normal expression. This decomposition is defined essentially by computing a linear model of diseased tissue expression data onto normal expression data. Equation (1) in Section 1 gives this decomposition, with the normal component $Nc.T$ being the least squares fit to normal tissue data, and the disease component $Dc.T$ the vector of residuals. However, in order to obtain a good approximation for normal expression data, we first reduce its dimension, using a modification of principal component analysis. Thus the *DSGA* decomposition in Equation (1) is based on a linear model to a reduced dimension approximation $\mathcal{N}$ of the normal expression data. Section 2.1 sketches the general setup for the method, Section 2.2 provides the details of dimension reduction and model fitting for the normal data and Section 2.3 uses data simulations to test the method of dimension reduction for the normal tissue data. Precise mathematical details found online: ***Computational Details Supplement*** Sections 1 and 2.

2.1 Data decomposition: the normal component and the disease component

We assume that microarray expression data has been collected for diseased tissue samples and for normal tissue samples. Each tissue sample data is a high dimensional vector in *array space* whose coordinates are genes:

- Diseased tissue microarray data: $T_1, T_2, \ldots, T_{\mathbf{S}}$
- Normal tissue microarray data: $N_1, N_2, \ldots, N_{\mathbf{R}}$

Note that we do *not* require that the number of normal tissue samples **R** be the same as the number of diseased samples **S**. In fact, ideally we would have for normal tissue microarray data a very large database (very large **R**) against which each diseased tissue data vector $T_{\mathbf{i}}$ would be decomposed. If the disease affects a specific organ, then all the normal tissue data should be collected from that particular organ.

Essentially, we first fit each data vector $T_{\mathbf{i}}$ from a diseased tissue sample to a linear model in the normal data $N_1, N_2, \ldots, N_{\mathbf{R}}$ defining its decomposition $T_{\mathbf{i}} = Nc.T_{\mathbf{i}} + Dc.T_{\mathbf{i}}$ with:

- $Nc.T_{\mathbf{i}}$ fit to the linear model: *Normal Component*
- $Dc.T_{\mathbf{i}}$ vector of residuals to linear model: *Disease Component*

It is a tacit assumption that normal tissue data will generally have intrinsic mathematical characteristics distinct from those of diseased tissue data. However, because of noise in the data, and because all tissue samples, including normal tissue, exhibit biological diversity, as the number of normal samples increases, so will the dimension of the normal data. Eventually, when the number of normal samples is larger than the number of genes, it is possible that the subspace generated by normal data will constitute the entire space, thereby making the residual vectors (disease component) for diseased tissue data into the **0**-vector. Thus, instead of using all the normal expression data, we first reduce the dimension of normal data as explained in Section 2.2 to obtain a better approximation $\mathcal{N}$ of a model for the normal expression space.

2.2 Estimation of the normal expression space $\mathcal{N}$

We use a modification of principal component analysis (*PCA*) to reduce the dimension of normal expression data: $N_1, N_2, \ldots, N_{\mathbf{R}}$. Although *PCA* is a natural method for reducing the dimension of this data, we have found that a modification of *PCA* works much better. The *flat* construction defined in Section 2.2.1 uses a series of linear model projections to give a cleaner estimate of the normal expression space $\mathcal{N}$. We then use *PCA* on data transformed with the *Flat* construction, rather than on the original normal tissue data. Data simulations in Section 2.3 show the utility of this construction in estimating $\mathcal{N}$.

We assume that normal tissue spans a subspace $\mathcal{N}$ of dimension k much smaller than the number R of normal tissues. Essentially, we assume that most normal expression lies in the space $\mathcal{N}$, and wish to recover from the normal data the space $\mathcal{N}$. We use a modification of the method originally defined by Wold (Eastment and Krzanowski, 1982; Krzanowski and Kline, 1995; Wold, 1978). When applying *DSGA* to microarray data in Section 3, dimension reduction to $\mathcal{N}$ was minimal, suggesting a need for more normal tissues. Despite this limitation, *DSGA* outperformed traditional methods.

2.2.1 Flat construction Starting with the normal tissue expression vectors $N_1, N_2, \ldots, N_{\mathbf{R}}$ we define new *flat* vectors: $\widehat{N}_1, \widehat{N}_2, \ldots, \widehat{N}_R$ by letting $\widehat{N}_i$ be the least squares fit of N_i to a linear model in all the other normal tissue arrays $N_1, N_2, \ldots, N_{i-1}, N_{i+1}, \ldots, N_R$. Roughly, working with the flat vectors is intended to reduce aspects of the data that are unique to each normal tissue expression vector $N_{\mathbf{i}}$, and are not (small) noise; rather they are (possibly large) biologically meaningful signal that is unique to N_i. Data simulations in Section 2.3 show that working with the *flat* vectors greatly improves our ability to recover the correct dimension reduction. We construct the matrix with columns the flat normal data $\widehat{\mathbf{N}} = [\widehat{N}_1\ \widehat{N}_2\ \ldots\ \widehat{N}_R]$.

2.2.2 The normal space $\mathcal{N}$ We wish to reduce the space generated by the *flat* normal vectors $\widehat{N}_1, \widehat{N}_2, \ldots, \widehat{N}_R$ to an appropriate principal component subspace. We use the method in (Wold, 1978). We compute for each $l < R$ the goodness of fit measure W for the Flat matrix $\widehat{\mathbf{N}}$.

$$W(l) \approx \left(\frac{\lambda_l^2}{\lambda_{l+1}^2 + \cdots + \lambda_R^2} \right) \frac{(n-l-1)(R-l)}{(n+R-2l)} \tag{2}$$

Here λ_i is the ith singular value of the *flat* normal data matrix $\widehat{\mathbf{N}}$, **R** is the number of columns (normal samples) and n is the number of rows (genes). Recall that λ_i essentially gives a measure of the amount of data in the ith direction, so that roughly, Wold's invariant $W(l)$ measures the ratio between the *smallest signal* (λ_l) and *all* noise (the subsequent singular values $\lambda_{l+1}, \ldots, \lambda_R$.)

We take L so that $W(l)$ spikes up for the value L, and construct the matrix $\widehat{\mathbf{N}}_L$ the top L-dimensional principal component approximation of the flat normal data matrix $\widehat{\mathbf{N}}$:

$$\widehat{\mathbf{N}}_L = U \cdot \Sigma_L \cdot V^t \tag{3}$$

where $\widehat{\mathbf{N}} = U \cdot \Sigma \cdot V^t$ is the singular value decomposition, and Σ_L is the diagonal matrix with the first L diagonal entries the same as the first L singular values for $\widehat{\mathbf{N}}$ and the rest of the entries 0.

The normal space $\mathcal{N}$ is the column space of $\widehat{\mathbf{N}}_L$.

2.3 Data simulation

We use data simulations to investigate the ability of *PCA* to detect an appropriate dimension reduction when combined with the *flat* construction. We compared *PCA* dimension reduction, with and without the *flat* transformation on simulated data. Roughly, we make the following assumptions about the normal data: we assume that (1) there is a gene expression signature common to all normal samples; (2) there is additional expression in normal data that varies continuously among the samples; (3) there is biological diversity providing uniqueness in global expression for each individual normal

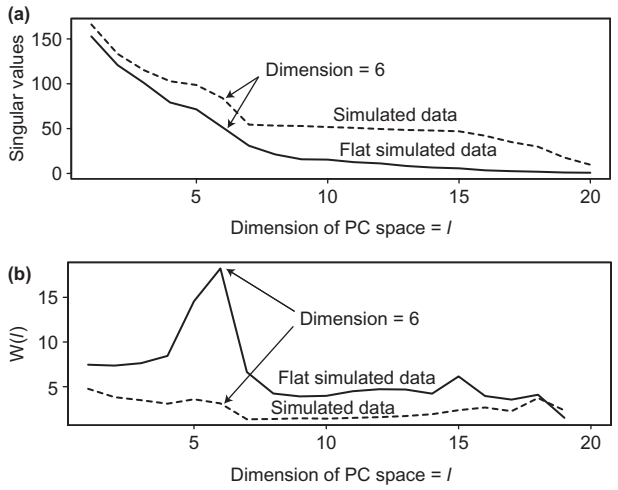

Fig. 2. Data simulation to determine correct dimension reduction for *PCA*. The correct dimension for the model should be 6. Graphs compare **(a)** singular values and **(b)** goodness of fit measure $W(l)$ versus the number of dimensions l. Graphs show simulated data and simulated *flat* data.

sample and (4) noise. Specifically, we assume the following model for the simulated *ith* normal array $N_i, i = 1, 2, \ldots, R$:

$$N_i = \left(\sum_{j=1}^{k} a_{ij} C_j\right) + b_i B + D_i + v_i \qquad (4)$$

Here B is a global feature common to all normal arrays, $C_1, \ldots, C_k$ span a virtual normal expression space of dimension k, D_i is data unique to the ith normal tissue, v_i is noise, a_{ij} and b_i are coefficients. The noise vector v_i is smaller than vectors B, C_j and D_i, and D_i represents real biology unique to the *ith* normal tissue.

For data simulations we assumed: 100 genes, $R = 20$, $k = 5$. We chose the collection of vectors $\{B, C_1, C_2, \ldots C_k\}$ to be mutually orthogonal. We wanted to recover the dimension: $k + 1 = 6$ from $W(l)$. We also wanted to obtain the space generated by $\{B, C_1, C_2, \ldots, C_k\}$ as the top $k + 1 = 6$ dimensional subspace by *PCA*. Details of this simulation are found in the *Computational Details Supplement Section* 2.1, along with additional simulations (Sections 2.2 and 2.3) investigating the effects of varying the parameters in Equation (4).

We computed the *flat* vectors $\widehat{N}_i$. We then computed the values for $W(l)$ for both collections: the original simulated normal tissue vectors $\{N_1, N_2, \ldots, N_R\}$ and the *flat* simulated data vectors: $\{\widehat{N}_1, \widehat{N}_2, \ldots, \widehat{N}_R\}$. Figure 2 shows a plot of the l versus $W(l)$ for both the original simulated data vectors and the *flat* simulated data vectors, showing that the values of W for the *flat* data can indeed recover the dimension $k + 1 = 6$. Moreover, when using the *flat* normal vectors $\{\widehat{N}_1, \widehat{N}_2, \ldots \widehat{N}_R\}$ the top six dimensional principal component subspace was indeed the subspace generated by the classes $\{C_1, C_2, \ldots, C_k\}$ together with the common signature B, but ignoring diversity vectors D_i and noise vectors v_i.

3 APPLICATION TO MICROARRAY DATA

We applied our analysis method to several cancer datasets. We first used these datasets to compare *DSGA* with other standard methods of analysis. Specifically, we used *PAM* to place tumors in different classes based on clinico-pathological characteristics, and compared error rates when the *PAM* analysis was performed on data that had been transformed in a variety of ways, including *DSGA*. In most cases, *DSGA* outperformed the other transformations. We then showed separately, by focusing on breast cancer, that *DSGA* has the potential to highlight novel biology, rather than merely identify, albeit with greater accuracy, already known properties. We first showed that the tumor class predictor genes identified by *PAM* in constructing the class shrunken centroids were largely different for *DSGA*-transformed data and for non-transformed data. We then went on to unravel a novel decomposition of the *ER* cluster in breast cancer.

3.1 Comparison of DSGA with other methods

We compared the ability of *PAM* (Tibshirani *et al.*, 2002) to make class predictions, for known clinico-pathological tumor distinctions. We used the following notation:
Gene expression cancer datasets were comprised of:

$$\text{Tumor arrays:} \quad T_1, T_2, \ldots, T_{\mathbf{S}}$$
$$\text{Normal array:} \quad N_1, N_2, \ldots, N_{\mathbf{R}}$$

Both sets of data consisted of log-transformed *cDNA* microarray expression data.

Data from tumor samples was then transformed in several different ways:

(1) *Traditional*, log.ratio data $\{T_i\}$

(2) *Zero-transformed* data $\{Zt.T_i\}$: the vector of gene means $\bar{N}$ of all the normal tissue data vectors was computed:

$$\bar{N} = mean(N_1, N_2, \ldots, N_{\mathbf{R}})$$

then the tumor data was transformed by subtracting $\bar{N}$ from each of the tumor data vectors:

$$Zt.T_i = T_i - \bar{N}$$

(3) *Paired normal-transformed* data $\{Npair.T_i\}$: when both tumor and normal tissue data is available for the same patient, the difference between tumor and normal data:

$$Npair.T_i = T_i - \ N_i$$

(4) *Disease components* from DSGA transformed data $\{Dc.T_i\}$.

We acknowledge that the number of patients for which paired samples—tumor and normal—was available was not very large. Moreover, only one of the cancer datasets had even a limited subcollection of paired data patients. Nevertheless, for the sake of completeness, we include in the supplement a comparison between *DSGA*-transformed and *paired normal-transformed data*, and note that in this case too, *DSGA* compares favorably with *paired normal transformation*. We note too that the paucity of paired tumor–normal data is often due to the difficulty in obtaining histologically normal tissue samples from a significant number of cancer patients.

3.1.1 Gastric cancer dataset Gene expression data from gastric cancer (Chen *et al.*, 2003) consisting of 89 tumor samples and 29 normal tissue samples was used. Of the 89 patients, 20 provided paired tissue samples: same-patient tumor and normal tissue. Data was retrieved as in (Chen *et al.*, 2003): either channel mean intensity over background

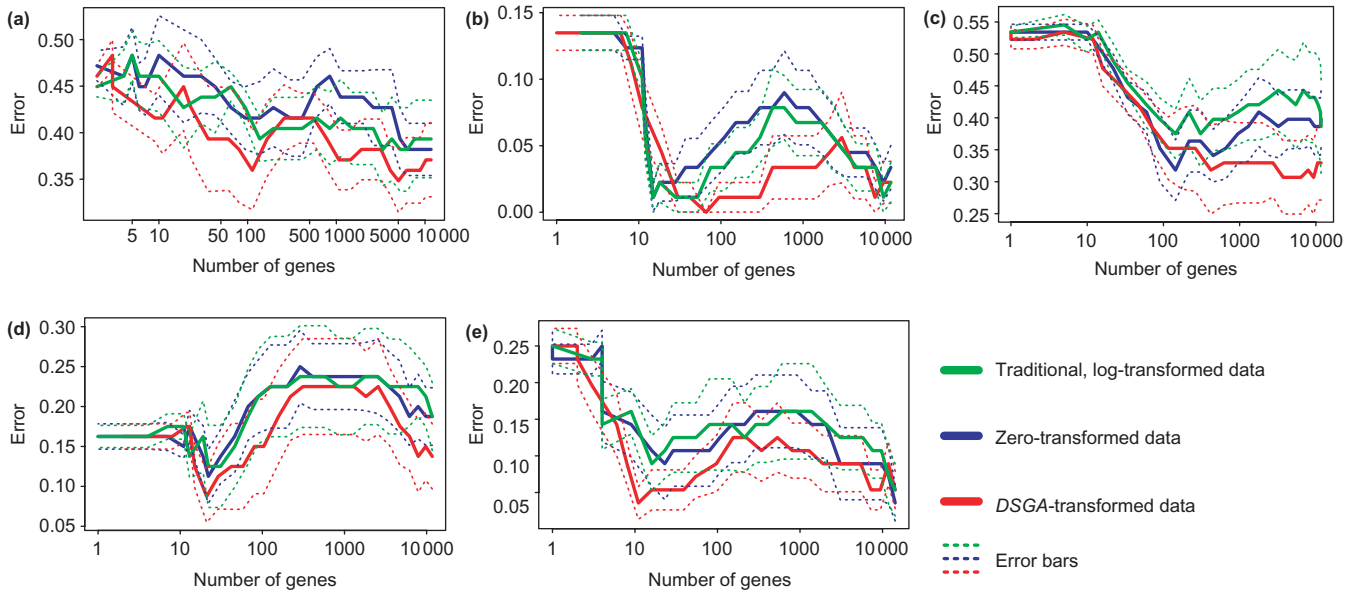

Fig. 3. Cross-validation error rates for class prediction by *PAM* for several known clinical distinctions in cancer datasets. Error rates for log-transformed data; zero-transformed data; and *DSGA*-transformed data are plotted on the same graphs (error bars—dashed). *DSGA*-transformed data outperforms traditional and zero-transformed data. **(a)** Gastric cancer: *Helicobacter Pylori* infection: positive versus negative; **(b)** Gastric cancer: *Epstein-Barr virus* infection: positive versus negative; **(c)** Gastric cancer: Tumor *site:* antrum versus body versus cardia; **(d)** Gastric cancer: Tumor *type:* diffuse versus intestinal; **(e)** Breast cancer: Tumor *type:* ductal versus classical lobular.

exceeded 3. Only clones with data for at least 80% of the samples were retained, and the remaining missing values were imputed using *KNN* algorithm (Troyanskaya *et al.*, 2001). Finally, data was collapsed (mean) by UniGene cluster, build 187 yielding 11,711 genes.

We considered four different clinico-pathological distinctions in the dataset, all known to associate with cancer progression and prognosis: (1) latent infection with *Helicobacter Pylori (HP),* (2) latent infection with *Epstein-Barr (EB)* virus, (3) tumor *site*: Antrum versus Body versus Cardia and (4) tumor *type*: Diffuse versus Intestinal. For each of these distinctions, we tested and compared the ability of *PAM* to distinguish tumors on the basis of data transformed in several ways. For the entire dataset, we compared the original log data $\{T_i\}$, the zero-transformed data $\{Zt.T_i\}$ and the disease components for the *DSGA*-transformed data $\{Dc.T_i\}$. The *DSGA* transformation was performed after the normal data was reduced from dimension 29 to 27 by the *flat* construction (Section 2.2) and *PCA*. Supplementary Figure 1S shows the plots for dimension reduction, and Figure 3 shows the *PAM* error rates.

For the smaller set of 20 tumors where data was available in pairs from the same patient: normal and tumor data, we compared disease components of *DSGA*-transformed data $\{Dc.T_i\}$ with the *paired normal-transformed* data $\{Npair.T_i = T_i - N_i\}$. Although the sample size was not very large, we include comparison of the performance with *PAM* of these two types of data transformations, both of which highlight a type of deviation in expression between tumor and normal tissue. As shown in Supplementary Figure 2S, while the ability to distinguish seems to be the same for both *HP* and *EB* class distinctions, *DSGA*-transformed data outperformed the paired normal data transformation for both tumor *site* and tumor *type* distinctions.

3.1.2 Breast cancer dataset Gene expression data from breast cancer (Zhao *et al.*, 2004) consisting of 63 primary tumor samples and 13 normal tissue samples was retrieved. Data was retrieved if either the spot regression correlation exceeded 0.6 or if both channels mean intensity over background exceeded 1.5. Only clones with data for at least 80% of the samples were retained, remaining missing values were imputed using *KNN* algorithm (Troyanskaya *et al.*, 2001). Finally, data was collapsed (mean) by UniGene cluster, build 187 yielding 14,237 genes. Most tumor samples, 57 of the original 63 tumors, were either *ductal* or *classical lobular* tumors. This distinction is known to be associated with a range of disease-related characteristics. We compared the original log ratio expression data $\{T_i\}$, the zero-transformed expression $\{Zt.T_i\}$, and the disease component for the *DSGA*-transformed data $\{Dc.T_i\}$. The *DSGA* transformation was performed after the normal data was reduced from dimension 13 to 12 by the *flat* construction (Section 2.2) and *PCA*. Supplementary Figure 3S shows the plots for dimension reduction, and Figure 3 shows the *PAM* error rates. The *DSGA*-transformed data outperformed both the original log ratio data and the zero-transformed data.

3.2 Comparison of *PAM* centroids for breast cancer

While the error rates for running *PAM* were improved with the *DSGA* transformation, we wanted to compare the predictor genes in the *PAM*-shrunken centroids. The diagram in Figure 4 shows the extent of overlap in the collection of predictor genes using the original log ratio data, zero-transformed data and *DSGA*-transformed data, and Supplementary Figure 4S shows in detail the shrunken centroids generated by *PAM*. While the zero-transformed gene list from *PAM* is a slight expansion of the original log ratio data gene list, with even the order of predictor genes being identical in both, all but 2 of the *DSGA* genes from *PAM* are different. This suggests that the underlying biology highlighted by the *DSGA* transformation may be quite different from that highlighted by log-transformed data, or the zero-transformed data.

3.3 Unraveling the estrogen receptor cluster in breast cancer

Cluster analysis can provide a wealth of information and often suggests putative biologically meaningful associations of genes (Eisen *et al.*, 1998). However, data transformations often change the

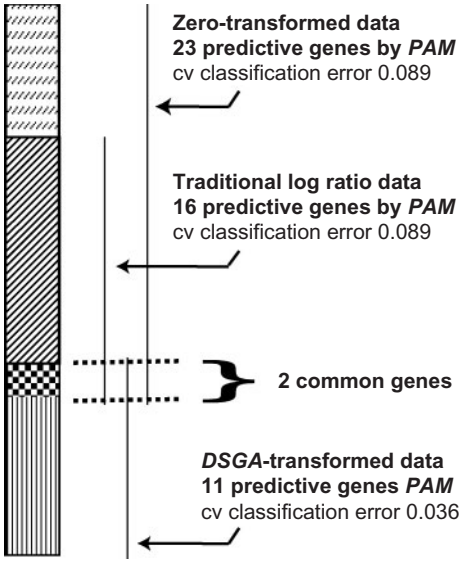

Fig. 4. Overlap of *predictive genes* in the *PAM*-shrunken centroids using traditional log-transformed data, zero-transformed data and *DSGA*-transformed data. Breast cancer dataset: tumor *type* class distinction *ductal* versus *classical lobular*. Zero-transformed data centroids were a slight expansion of the log-transformed data centroids. By contrast, these had only two genes in common with the centroids produced by *DSGA* data.

mathematical associations between genes, and consequently they can drastically affect the clustering. We investigated the effect of the *DSGA* transformation on clustering the breast cancer dataset, and compared it to clustering the same data, using traditional, gene mean-centered, log—transformed data. Clustered data was viewed using Java Treeview (Saldanha, 2004). Specifically we focused on a specific cluster of genes, known to be important in breast cancer—the *ER* cluster. *ER*, and generally hormone receptor status, is known to be profoundly involved in the pathology of breast cancer. The involvement of *ER* is so fundamental that a multitude of variables are associated with *ER* status in breast cancer: age, time to metastasis, overall survival and response to therapy, and there is strong evidence that *ER* coregulation, and *GATA3* coexpression constitute strong outcome predictors for a large class of breast cancers—luminal breast cancers (Oh *et al.*, 2006). Thus understanding *ER* coregulation is a fundamental step in unraveling the underlying biology of various types of breast cancer.

One important advantage to using *DSGA*-transformed data is that all expression is relative to a biologically meaningful standard: expression levels in normal tissue. For log ratio expression data, Pearson correlation of gene mean-centered data identifies two genes as highly similar (correlated) as long as their expression relative to the mean is similar. This can occur even if one gene is consistently over-expressing relative to normal tissue, and the other is consistently under-expressing. For *DSGA*-transformed data, we retain expression relative to normal tissue levels, and the distinction between genes that are over- and under-expressing relative to normal tissue can be easily identified by using uncentered correlation.

To investigate the effect on clustering of the *DSGA* decomposition, we considered the breast cancer dataset consisting of all 63 primary tumor samples: ductal, classical lobular, solid, trabecular alveolar lobular and classic trabecular lobular tumors. UniGene-collapsed data was further reduced by testing (1) *Deviation from a normal tissue expression* null hypothesis, and (2) *Deviation from mean* null hypothesis, as we now explain.

(1) Deviation from normal expression null hypothesis A leave-one-out step was performed on the normal dataset, by computing disease component of each normal tissue expression vector N_i using as normal data the *flat* normal dataset of all normal array data, excluding N_i. The *PCA* dimension reduction used was the same as that obtained for the original normal dataset: $dim = 12$. This produced disease components for each normal tissue expression vector: $Dc.N_1, Dc.N_2, \ldots, Dc.N_k$. For each gene G, the 95*th* percentile Q_G95 of absolute value of leave-one-out residuals was computed, as was the 99*th* percentile of these for all genes: $Q99$. This defined a filter bound for each gene *G Normal Filter* to be the greater of Q_G95 and Q99.*DSGA* was performed on the tumor data, and for each gene in the disease components of tumors the 5*th* and 95*th* percentiles were computed for the entire set of tumors. Genes were retained if the larger in absolute value of the 5*th* and 95*th* percentiles for the genes exceeded the filter *Normal Filter*. This step reduced the total number of genes to 1610.

(2) Deviation from mean tumor expression null hypothesis For each gene G retained above, the difference between the 95*th* and 5*th* percentiles was computed, and genes were retained if this exceeded the top 45*th* percentile of all such deviations for the retained genes. This step reduced the total number of genes to 885. Data was then clustered as follows:

- Arrays were clustered by *Pearson correlation*: disease components of *DSGA* tumor data $\{Dc.T_i\}$ were gene mean-centered.
- Genes were clustered by *uncentered correlation* of disease components of *DSGA* tumor data $\{Dc.T_i\}$ (**not** mean-centered.)
- The heatmap for clustered data shows the disease component values ($Dc.T_i$) for each gene. Thus **up** and **down** regulation in the heatmap indicates **up** and **down** regulation relative to normal tissue levels.

Additionally, a **0** array vector, a *virtual normal array*, was included in the *DSGA* decomposed dataset of tumors, prior to clustering, thereby providing additional information for comparing tumor data with normal expression.

In order to compare with clustering on traditional, log-transformed data, unsupervised hierarchical clustering was performed on the same dataset of 63 tumors, with three normal tissue arrays. The dataset was gene and array mean-centered, and genes were retained if they deviated from the mean by at least $log_2(3)$ on at least three arrays. This reduced the number of genes to 2287. Hierarchical clustering was then performed on this reduced dataset. Figure 5a shows side by side the two heatmaps resulting from clustering the traditional log-transformed and the *DSGA*-transformed datasets. Many distinctions between the two analyses ensue, but our focus is the *estrogen receptor* co-expressing genes: the *ER*-cluster.

Interestingly, the *DSGA* decomposition causes a splitting of the traditional *ER* cluster into at least three distinct subclusters: a proper *ER*-associated cluster, a *Forkhead box A1 (FOXA1) and GATA3* associated cluster and a *Signal peptide, CUB domain, EGF-like 2 (SCUBE2)* associated cluster. These three clusters show coherent expression in corresponding clusters of tumors, known to be of distinct cancer phenotype: (1) tumors showing low expression of *ER* and *ERBB2* or HER2/neu; (2) tumors showing low *ER* expression but over-expression of *ERBB2* or HER2/neu; (3) lobular tumors and (4) ductal *ER* positive tumors. We emphasize that, since data is *DSGA* transformed, positive and negative status are indeed relative to normal expression, rather than to the mean expression of the group of tumors included in the study. Figure 5b shows the detail of the *DSGA* decomposition of the *ER* cluster into these three clusters, along

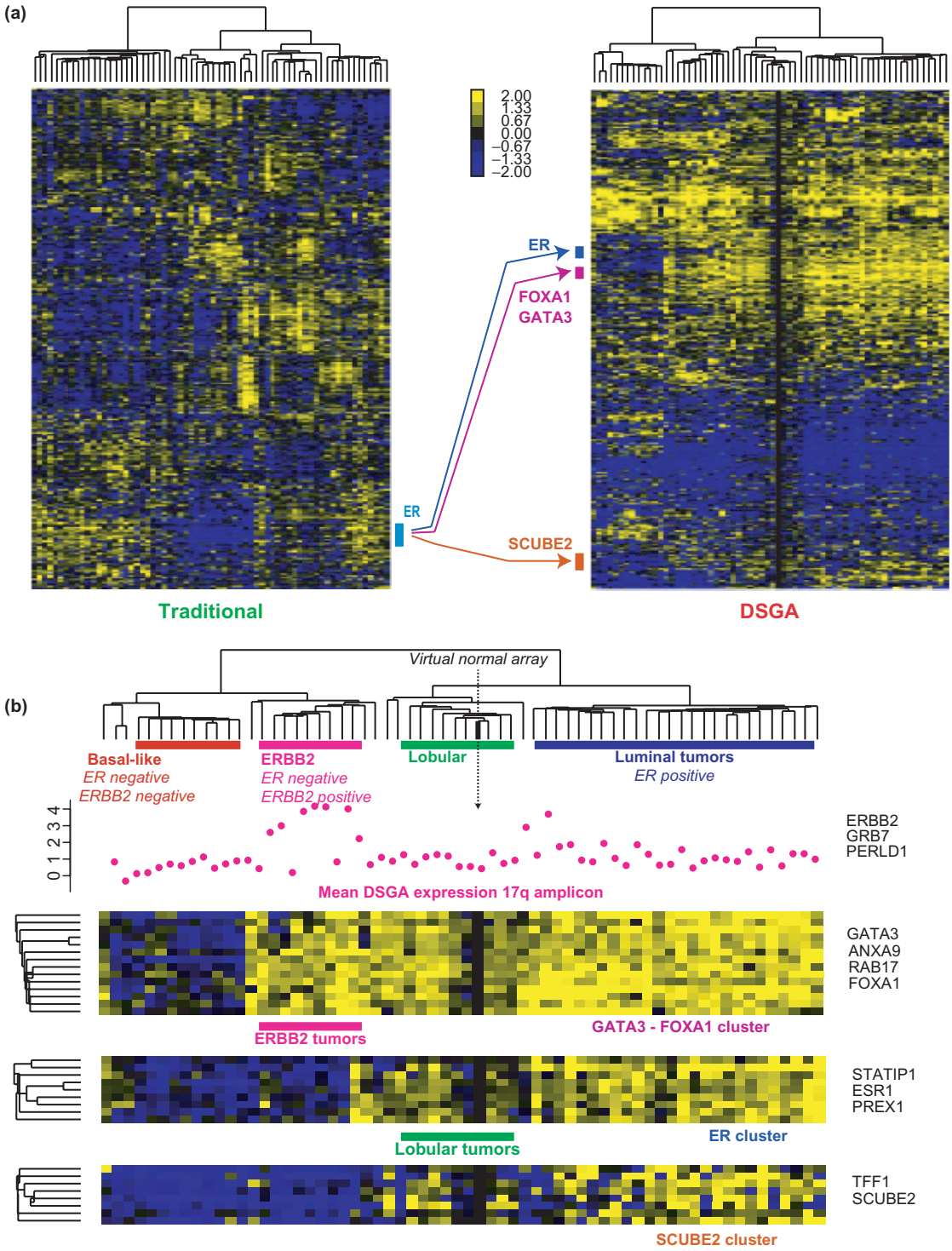

Fig. 5. Decomposition of the *ER* gene cluster as a consequence of using the *DSGA* transformation on a breast cancer dataset. **(a)** Comparison of global heatmaps showing hierarchical clustering on the data. The position of the traditional *ER* cluster using log-transformed data, and its splitting into three separate smaller clusters in the *DSGA*-transformed heatmap are shown. TRADITIONAL heatmap data values are gene and array mean-centered. *DSGA* heatmap data values are the *disease components* of the data (deviation from normal expression.) The traditional *ER* cluster unravels into three clusters in the *DSGA* analysis: *ER* cluster; *FOXA1–GATA3* cluster; and *SCUBE2* cluster. **(b)** Close-up view of the three *DSGA* gene clusters, together with a comparison to the mean *DSGA*-expression levels of the *17q amplicon* containing the *ERBB2* gene. The distinction between the *GATA3–FOXA1* cluster and the *ER* cluster occurs primarily along tumors that are *ER* negative, *ERBB2* overexpressing. The distinction between the *ER* cluster and the *SCUBE2* cluster occurs primarily along lobular tumors, and *ER* positive, *ERBB2* overexpressing tumors.

with the expression level of the ERBB2/17q amplicon. It is evident that the distinction between the *FOXA1*/*GATA* 3 cluster and the *ER* cluster occurs primarily within the *ER*-negative *ERBB2* or HER2/neu-overexpressing tumors. The distinction between the *ER* cluster and the *SCUBE2* cluster appears to be mostly within the lobular tumors, as well as the *ER*-positive *ERBB2* or HER2/neu-overexpressing tumors.

While we acknowledge the need for an in-depth extensive analysis of the split in the traditional *ER* cluster, this exploratory analysis suggests that further investigation in the differential disruption of co-expression for these genes may highlight distinctions in the underlying biology of what are known to be different molecular subtypes of breast cancer.

4 DISCUSSION

DSGA highlights a series of issues that merit further investigation. Our understanding of disease necessitates extensive understanding of normal phenotypes to ensure that the characteristics we study are indeed aberrant and clearly deviate from the realm of healthy phenotype. Extensive normal data would thus expand our understanding of disease. The paucity of available normal expression data may explain the phenomenon observed in Section 3 where dimension reduction for normal data was minimal for both gastric and breast cancer datasets. Additional normal data would also allow investigating methods to assess the relative benefit of including additional normal data or tumor data in studying the disease. Supplementary Figure 11S provides a comparison using the gastric cancer data. Finally, extensive normal tissue data provides an opportunity to investigate optimal models for normal expression.

DSGA is intended to address a series of biological characteristics of diseased tissue expression and normal tissue expression:

(1) Our definition of *disease* is the **deviation** of expression from the *normal* or healthy state; thus rather than merely identifying variables (genes) whose expression is significantly distinct in diseased versus normal tissue and working with the original diseased tissue data along these significant genes, we first decompose the original data T into normal-like expression $Nc.T$ and deviation $Dc.T$ from normal-like expression. The disease is then defined to be the difference ($Dc.T$) between diseased tissue expression and normal-like expression.

(2) Our model $\mathcal{N}$ for the *normal* state incorporates some of the biological diversity inherent in normal tissue. This diversity stems from a multitude of sources: normal expression fluctuates in response to different conditions; normal tissue of distinct individuals can vary extensively; normal tissue is composed of a many distinct cell types with distinct expression patterns. The space $\mathcal{N}$ consists of *linear* combinations of normal data, thus providing a continuum of virtual normal expression vectors representing a range of combinations of these varied normal phenotypes, including a range of cell type mixtures.

(3) We do not require that each patient provide a normal tissue sample. This is important from a practical viewpoint, since for many patients the entire organ is visibly altered by the presence of the disease thereby making it impossible to obtain such samples.

(4) Each diseased tissue sample is analyzed and decomposed along the normal tissue null hypothesis alone, *without* reference to any other diseased tissues in the study. The disease component of each individual diseased tissue is obtained from the original array data vector T and the entire normal state model $\mathcal{N}$. In particular, the disease component $Dc.T$ is independent of the particular collection of diseased tissues included in a study. This is not the case, for example, when data is transformed by gene mean-centering, since the mean of each gene is determined by the values for the entire collection of samples.

We saw that *DSGA* outperforms other methods for class prediction where the classes were defined in terms of clinicopathology known to be relevant to outcome of disease.

5 CONCLUSION

We introduced a method for analysis of microarray data that highlights and separates aberrant expression in diseased tissue in order to understand the underlying biology of the pathologic process. The method first uses linear models (flat construction) and principal component analysis to construct a normal expression null hypothesis space $\mathcal{N}$. The diseased tissue expression data is then decomposed into two orthogonal components: the normal component best mimics normal expression in terms of linear models, and the disease component measures the deviation from the normal expression null hypothesis.

ACKNOWLEDGEMENTS

The authors are grateful to X. Chen, Y. Ji and H. Zhao for the cancer data, and to P. Norvig, G. Sherlock, G. Walther and A. Whittemore for illuminating conversations. This work was supported by the NIH-NHGRI: K01 grant HG00030 (M.N); NIH-NCI: U01 grant CA085129 (S.S.J); California Breast Cancer Research Program of the University of California, Grant Number 10EB-1086 (S.S.J. and M.N.); NSF grant DMS–9971405 (R.T.); and NIH contract N01-HV-28183 (R.T.).

Conflicts of Interest: none declared.

REFERENCES

Alon,U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Bair,E. *et al.* (2006) Prediction by supervised principal components. *J. Am. Stat. Assoc.*, **199**, 133–154.

Boer,J.M. *et al.* (2001) Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. *Genome Res.*, **11**, 1861–1870.

Chen,X. *et al.* (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, **13**, 1929–1239.

Chen,X. *et al.* (2003) Variation in gene expression patterns in human gastric cancers. *Mol. Biol. Cell*, **14**, 3208–3215.

Creighton,C.J. *et al*. (2006) Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. *Genome Biol*., **7**, R28.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 78–87.

Dudoit,S. *et al*. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray expreriments. *Stat. Sin*., **12**, 111–139.

Eastment,H.T. and Krzanowski,W.J. (1982) Cross-validatory choice of the number of components from a Principal Components Analysis. *Technometrics*, **24**, 73–77.

Eisen,M. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Foekens,J.A. *et al*. (2006) Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J. Clin. Oncol*., **24**, 1665–71.

Ghosh,D. (2004) Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics*, **20**, 1663–1669.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gruvberger,S. *et al*. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res*., **61**, 5979–5984.

Hastie,T. *et al*. (2000) Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*., **1**, 8418–8423.

Innes,H.E. *et al*. (2006) Significance of the metastasis-inducing protein AGR2 for outcome in hormonally treated breast cancer patients. *Br. J. Cancer*, **94**, 1057–1065.

Krzanowski,W.J. and Kline,P. (1995) Cross-validation for choosing the number of important components in principal components analysis. *Multivariate Behav. Res*., **30**, 149–165.

Laganiere,J. *et al*. (2005) Location analysis of estrogen receptor α target promoters reveals that FOXA1 defines a domain of the estrogen response. *Proc. Natl Acad. Sci. USA*, **102**, 11651–11656.

Munagala,K. *et al*. (2004) Cancer characterization and feature set extraction by discriminative margin clustering. *BMC Bioinformatics*, **5**, 21.

Oh,D.S. *et al*. (2006) Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *J. Clin. Oncol*., **24**, 1656–1664.

Paik,S. *et al*. (2004) A multigene assay to predict recurrence of tamisifen-treated, node-negative breast cancer. *N. Engl. J. Med*., **351**, 2817–2826.

Perou,C.M. *et al*. (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Saldanha,A. J. (2004) Java Treeview–extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.

Sørlie,T. *et al*. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Sørlie,T. *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.

Stephanopoulos,G. *et al*. (2002) Mapping physiological states from microarray expression measurements. *Bioinformatics*, **18**, 1054–1063.

Tibshirani,R. *et al*. (2002) Diagnostic of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Tronyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Usary,J. *et al*. (2004) Mutation of GATA3 in human breast tumors. *Oncogene*, **23**, 7669–7678.

van de Vijver,M.J. *et al*. (2002) A gene expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med*., **347**, 1999–2009.

Wang,Y. *et al*. (2005) Gene Expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*., **365**, 671–679.

Weigelt,B. *et al*. (2005) Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Res*., **65**, 9155–9158.

Weinstein,J. *et al*. (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science*, **275**, 343–349.

Wold,S. *et al*. (1978) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**, 397–405.

Yang,F. *et al*. (2006) Laser microdissection and microarray analysis of breast tumors reveal ER-alpha related genes and pathways. *Oncogene*, **25**, 1413–1419.

Zhao,H. *et al*. (2004) Different gene expression patterns in invasive lobular and ductal carcinomas ofthe breast. *Mol. Biol. Cell*., **15**, 2523–2536.