Gene expression

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

Monica Nicolau\textsuperscript{1,2}, Robert Tibshirani\textsuperscript{3,4}, Anne-Lise Børresen-Dale\textsuperscript{5,6} and Stefanie S. Jeffrey\textsuperscript{1,}\textsuperscript{*}  

\textsuperscript{1}Department of Surgery, Stanford University School of Medicine, \textsuperscript{2}Department of Mathematics, \textsuperscript{3}Department of Health, Research & Policy, \textsuperscript{4}Department of Statistics, Stanford University, \textsuperscript{5}Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Center and \textsuperscript{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

\textbf{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, \textit{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: ssi@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 \textit{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 \textit{et al.}, 1999; Dudoit Fridlyand \textit{et al.}, 2002; Eisen \textit{et al.}, 1998; Golub \textit{et al.}, 1999; Hastie \textit{et al.}, 2000; Weinstein \textit{et al.}, 1997), genes whose expression is stable among different samples from the same patient (Sorlie \textit{et al.}, 2003; Weigelt \textit{et al.}, 2005), genes whose expression levels show a strong association with various clinico-pathologic characteristics (Bair \textit{et al.}, 2006; Dudoit Yang \textit{et al.}, 2002; Tusher \textit{et al.}, 2001; Vijver \textit{et al.}, 2002; Weigelt \textit{et al.}, 2005), and various methods that identify genes whose expression most significantly distinguish diseased and normal tissues: (Alon \textit{et al.}, 1999; Boer \textit{et al.}, 2001; Chen \textit{et al.}, 2002; Ghosh \textit{et al.}, 2004; Munagala \textit{et al.}, 2004; Stephanopoulos \textit{et al.}, 2002).

In this article we introduce a novel method of data analysis: \textit{disease-specific genomic analysis (DSGA)} that employs comparison to normal expression to extract data most closely associated with the disease. Specifically, \textit{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, \textit{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 \textit{after} the data has been transformed by \textit{DSGA}. This method defines the mathematical model for \textit{normal} expression to be a linear subspace derived from normal tissue expression data; it defines \textit{disease-specific} expression to be the deviation of expression in diseased tissue from this subspace, where \textit{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 \textit{normal component} \( N_e.T \) is the least squares fit of \( T \) to a linear model in \( \mathcal{N} \), and the \textit{disease component} \( D(T) \) is the vector of residuals from the fit.
to this linear model. The two vectors $Nc.T$ and $Dc.T$ are perpendicular, and satisfy:

$$T = Nc.T + Dc.T$$  \hspace{1cm} (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., 2000; 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; Sorlie et al., 2001; Sorlie 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 reduced dimension approximation $V'$ 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: \( T_1, T_2, \ldots, T_S \)
- Normal tissue microarray data: \( N_1, N_2, \ldots, N_R \)

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 \( T_i \) 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 \( T_i \) from a diseased tissue sample to a linear model in the normal data \( N_1, N_2, \ldots, N_R \) defining its decomposition \( T_i = Nc.T_i + Dc.T_i \) with:

- \( Nc.T_i \) fit to the linear model: Normal Component
- \( Dc.T_i \) 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 \( \theta \)-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: \( N_1, N_2, \ldots, N_R \). 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 \( N_1, N_2, \ldots, N_R \) we define new flat vectors: \( \tilde{N}_1, \tilde{N}_2, \ldots, \tilde{N}_R \) by letting \( \tilde{N}_i \) be the least squares fit of \( N_i \) to a linear model in all the other normal tissue arrays \( N_1, N_2, \ldots, \hat{N}_{i-1}, N_{i+1}, \ldots, N_R \). Roughly, working with the flat vectors is intended to reduce aspects of the data that are unique to each normal tissue expression vector \( N_i \), 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 \( \tilde{N} = [\tilde{N}_1 \tilde{N}_2 \ldots \tilde{N}_R] \).

2.2.2 The normal space \( \mathcal{N} \) We wish to reduce the space generated by the flat normal vectors \( \tilde{N}_1, \tilde{N}_2, \ldots, \tilde{N}_R \) 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 \( \tilde{N} \):

\[
W(l) \approx \left( \frac{\lambda_i}{\lambda_i + \lambda_j + \cdots + \lambda_R} \right) \frac{(n-l-1)(R-l)}{(n+R-2l)}
\]

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

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

\[
\tilde{N}_L = U \cdot \Sigma_L \cdot V' \tag{3}
\]

where \( \tilde{N} = U \cdot \Sigma \cdot V' \) is the singular value decomposition, and \( \Sigma_L \) is the diagonal matrix with the first \( L \) diagonal entries the same as the first \( L \) singular values for \( \tilde{N} \) and the rest of the entries 0. The normal space \( \mathcal{N} \) is the column space of \( \tilde{N}_L \).

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
Computational Details Supplement Section

found in the

Investigating the effects of varying the parameters in along with additional simulations (Sections 2.2 and 2.3) the values for \( f \) generated by \( \text{mutually orthogonal} \). We wanted to recover the dimension: \( \text{f} \text{Ck} \text{D} \text{i} \text{R} \text{C} \text{k} \) \( \text{f} \text{C} \text{k} \text{D} \text{i} \text{R} \text{C} \text{k} \). We also wanted to obtain the space \( \text{flat} \text{f} \text{C} \text{R} \text{C} \text{k} \text{D} \text{R} \text{C} \text{k} \), \( \text{flat} \text{f} \text{C} \text{k} \text{D} \text{R} \text{C} \text{k} \), and \( \text{flat} \text{f} \text{C} \text{k} \text{D} \text{R} \text{C} \text{k} \) simulated data vectors: \( \text{flat} \text{f} \text{C} \text{k} \text{D} \text{R} \text{C} \text{k} \) showing \( \text{flat} \text{f} \text{C} \text{k} \text{D} \text{i} \text{R} \text{C} \text{k} \) and (4) noise. Specifically, we assume the following model for the simulated \( \text{ith} \) normal array \( N_i, i = 1, 2, \ldots, R \):

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

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 \( \text{ith} \) 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 \( \text{ith} \) 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 \( \text{flat} \) vectors \( \tilde{N}_i \). 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 \( \text{flat} \) simulated data vectors: \( \{\tilde{N}_1, \tilde{N}_2, \ldots, \tilde{N}_R\} \). Figure 2 shows a plot of the \( l \) versus \( W(l) \) for both the original simulated data vectors and the \( \text{flat} \) simulated data vectors, showing that the values of \( W \) for the \( \text{flat} \) data can indeed recover the dimension \( k + 1 = 6 \). Moreover, when using the \( \text{flat} \) normal vectors \( \{\tilde{N}_1, \tilde{N}_2, \ldots, \tilde{N}_R\} \) 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:

- Tumor arrays: \( T_1, T_2, \ldots, T_9 \)
- Normal array: \( N_1, N_2, \ldots, N_R \)

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 \( T_l \)**
2. **Zero-transformed data \( Z_l; T_l \)**: the vector of gene means \( \tilde{N} \) of all the normal tissue data vectors was computed:
   \[
   \tilde{N} = \text{mean}(N_1, N_2, \ldots, N_R)
   \]
   then the tumor data was transformed by subtracting \( \tilde{N} \) from each of the tumor data vectors:
   \[
   Z_l; T_l = T_l - \tilde{N}
   \]
3. **Paired normal-transformed data \( N_{pair}; T_l \)**: when both tumor and normal tissue data is available for the same patient, the difference between tumor and normal data:
   \[
   N_{pair}; T_l = T_l - N_l
   \]
4. **Disease components from DSGA transformed data \( Dc; T_l \)**

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 historically 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—tumor and normal tissue. Data was retrieved as in (Chen et al., 2003): either channel mean intensity over background
least 80% of the samples were retained, remaining missing values were imputed using KNNo
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 $\{Z_i,T_i\}$ 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 $\{N_{pair},T_i = T_i - N_i\}$. 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 KNNo 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 $\{Z_i,T_i\}$, 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
DSGA-transformed data
11 predictive genes by PAM
cv classification error 0.036

2 common genes

Traditional log ratio data
16 predictive genes by PAM
cv classification error 0.089

Zero-transformed data
23 predictive genes by PAM
cv classification error 0.089

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 normal expression null hypothesis, and (2) Deviation from mean expression 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: \(Dc_N1, Dc_N2, \ldots, Dc_N6\). For each gene \(G\), the 95th percentile \(Q_{95}\) of absolute value of leave-one-out residuals was computed, as was the 99th percentile of these for all genes: \(Q_{99}\). This defined a filter bound for each gene \(G\) Normal Filter to be the greater of \(Q_{95}\) and \(Q_{99}\). DSGA was performed on the tumor data, and for each gene in the disease components of tumors the 5th and 95th percentiles were computed for the entire set of tumors. Genes were retained if the larger in absolute value of the 5th and 95th percentiles for the genes exceeded the filter Normal Filter. This step reduced the total number of genes to 1610.

Data was then clustered as follows:

- Arrays were clustered by Pearson correlation: disease components of DSGA tumor data \(\{Dc_T\}\) were gene mean-centered.
- Genes were clustered by uncentered correlation of disease components of DSGA tumor data \(\{Dc_T\}\) (not mean-centered.)
- The heatmap for clustered data shows the disease component values \(\{Dc_T\}\) 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 \(log(3)\) 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 into normal-like expression $N_c.T$ and deviation $D_c.T$ from normal-like expression. The disease is then defined to be the difference ($D_c.T$) between diseased tissue expression and normal-like expression.

2. Our model $N$ 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 $N$ 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 $N$. 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 clinico-pathology 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 $N$. 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


Disease-specific genomic analysis (DSGA)


