Validation of Gene Expression Profiles in Genomic Data through Complementary Use of Cluster Analysis and PCA-Related Biplots

Niccolò Bassani*, Federico Ambrogi#, Danila Coradini#, Patrizia Boracchi# and Elia Biganzoli#

Department of Clinical Sciences and Community Health, University of Milan, via Vanzetti 5, 20133 Milano (MI), Itlay

Abstract: High-throughput genomic assays are used in molecular biology to explore patterns of joint expression of thousands of genes.

These methodologies had relevant developments in the last decade, and concurrently there was a need for appropriate methods for analyzing the massive data generated.

Identifying sets of genes and samples characterized by similar values of expression and validating these results are two critical issues related to these investigations because of their clinical implication. From a statistical perspective, unsupervised class discovery methods like Cluster Analysis are generally adopted.

However, the use of Cluster Analysis mainly relies on the use of hierarchical techniques without considering possible use of other methods. This is partially due to software availability and to easiness of representation of results through a heatmap, which allows to simultaneously visualize clusterization of genes and samples on the same graphical device. One drawback of this strategy is that clusters' stability is often neglected, thus leading to over-interpretation of results.

Moreover, validation of results using external datasets is still subject of discussion, since it is well known that batch effects may condition gene expression results even after normalization.

In this paper we compared several clustering algorithms (hierarchical, k-means, model-based, Affinity Propagation) and stability indices to discover common patterns of expression and to assess clustering reliability, and propose a rank-based passive projection of Principal Components for validation purposes.

Results from a study involving 23 tumor cell lines and 76 genes related to a specific biological pathway and derived from a publicly available dataset, are presented.

Keywords: Microarrays, cluster stability, multivariate visualization, Principal Components Analysis, cell polarity.

INTRODUCTION

DNA microarrays, also known as gene chip, are a multiplex technology which dates back to almost twenty years ago. Since then, it has known a relevant development becoming a standard technique for genomic analysis. In parallel, there has been a considerable effort to develop adequate statistical methods for dealing with this kind of data [1].

Many standard techniques for multivariate data, such as cluster analysis [2] and principal components [3-5], have been used at length to analyze gene expression datasets.

Cluster analysis, for instance, is the method of choice to discover groups of genes or samples with similar levels of expression and includes a countless series of algorithms to be used. Yet, often only one of these is used and the results are not further investigated by using some other methods [6, 7]. Additionally, stability of clustering results should be properly evaluated to assess robustness of the discovered groups [8].

Once the clusters have been defined, it is of common interest to investigate association between samples and genes, evaluating whether some panel of genes, possibly belonging to a common biological pathway, characterize a specific cluster. To do this, Principal Components Analysis (PCA) is a very powerful technique, allowing for a graphical representation of such an association using the biplot, a bivariate visualization of multivariate data introduced by Gabriel [9]. By looking at this graph, where the first two principal components are plotted, it is possible not only to explore the association between variables (genes) and observations (samples), but one can also evaluate the relationship among the variables themselves, gaining a considerable amount of information about gene expression levels.

Results deriving from these analysis need however to be validated, a task which is not trivial, as it has been reported since over ten years that "batch effects" need to be seriously taken into account when comparing results from different studies [10].

*Address corresponding to this author at the Department of Clinical Sciences and Community Health, University of Milan, via Vanzetti 5, 20133 Milano (MI), Itlay; Tel: 02 23902065; Fax: 02 50320866; E-mail: niccolo.bassani@unimi.it

#Co-Authors: federico.ambrogi@unimi.it, danila.coradini@yahoo.it, patrizia.boracchi@unimi.it, elia.biganzoli@unimi.it

E-ISSN: 1929-6029/12

Aim of the present study is thus to describe how cluster analysis and PCA-related biplots can be complementarily used to extract reliable information on samples classification and on the association between samples and genes and to validate experimental results.

Specifically, four clustering algorithms have been compared (hierarhical, k-means, model-based and Affinity Propagation) and stability of the results has been assessed using different indices. We addressed the issue of validation through a rank-based passive projection of the validation samples on the biplot of the experimental samples, in order to circumvent the "study effect" and to compare gene expression profiles across different experiments.

As a motivating example, we describe results of a study performed on a set of epithelial tumor cell lines (derived from the NCI60 dataset [11,12]) to evaluate patterns of expression of a panel of genes involved in the process of cell polarity. The biological relevance of the study lies in the fact that a strong correlation between malignancy and loss of epithelial organization has been histologically documented for almost types of tumor deriving from epithelial cells [13-16]. In addition, disruption of cell-cell junctions per se has been found to promote the development of some cancers [17,18]. Therefore, understanding the molecular mechanisms that regulate tissue organization and how such mechanisms are disrupted during neoplastic transformation, could provide important and useful insights to be exploited for diagnostic and therapeutic purposes.

MATERIALS AND METHODS

Clustering Algorithms

Hierarchical

Hierarchical clustering has been widely used in microarray data, starting from the seminal paper of Eisen *et al.* [19], mainly because of the possibility to graphically visualize clustering results by means of a heatmap, a plot where gene expression values are represented as a matrix of small coloured squares. In microarray studies, heatmaps are often on a green-red scale, where green stands for low expression values and red stands for high expression values. Additionally, genes and samples are re-ordered according to hierarchical clustering results, and related dendrograms are shown on the heatmap itself.

In this paper we used agglomerative hierarchical clustering with average linkage, using 1 – Pearson's ρ^2 as a distance measure, where ρ is defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

K-Means

The k-means algorithm was considered to have a standard non-hierarchical method for data clustering. The algorithm was initialized using the means of the clusters identified by the hierarchical clustering method described before. Choice of the number of clusters for both hierarchical and k-means algorithms was based on graphical visualization of mean silhouette value [2].

Affinity-Propagation

Affinity-Propagation (AP) [20] is a method based on graphical models which aims at searching for some centroids, named *exemplars*, into the set of data points possibly by starting from a pre-defined *preference value* for each object to be clustered or using the same for all objects.

As first step of the algorithm a similarity measure for each pair of subjects *(i,j)*, indicated as *s(i,j)*, is defined: this value suggests how much subject *j* is likely to be the centroid for subject *i*; then each subject is assigned a preference value *s(i,i)* which measures how much a point is suitable for being an exemplar. The core of the method is the exchange of messages between all pairs *(i,j)*; the first type of message, named *responsibility*, is defined as

$$r(i,j) = s(i,j) - \max_{j^{'}:j \neq j^{'}} \left\{ a\left(i,j^{'}\right) + s\left(i,j^{'}\right) \right\}$$

and represents how much *j* is suitable to be a centroid for *i*, considering all other potential exemplars for *i*. The second type of message, named *availability*, is defined as

$$a(i,j) = \min\left\{0, r(j,j) + \sum_{i^{'}:i^{'} \notin \{i,j\}} \max\left\{0, r\left(i^{'},j\right)\right\}\right\}$$

and indicates how much appropriate it would be for subject *i* to choose *j* as its centroid. In this equation *r(j,j)* is called "self-responsibility", and reflects evidence for *j* being an exemplar and against its grouping with a cluster identified by another centroid. At the first iteration the value for *a(i,j)* is set to zero, and all the pair-wise responsibilities are computed; then, starting from current values of *r(i,j)*, *a(i,j)* can be computed. At each iteration these messages can be combined to identify the exemplars for each of the subjects: algorithm stops at a determined number of iterations, or when a stable clustering result is found.

A preference value is associated uniquely with a clustering solution, but the same solution can be associated to different values of *s(i,i)*. By plotting on the x-axis the various preference values and on the y-axis the number of clusters associated it is possible to evaluate the plateaus in corrispondence of specific values of *k* to choose the number of clusters. By using this method, Soria *et al.* [21] evaluated gene profiles of several cases of breast cancer using AP, and obtaining results consistent with previous findings, but with an indication about the number of clusters.

In this paper we used Pearson's ρ as similarity measure, and the median of all pair-wise correlations between samples as the starting preference value *s(i,i)* for all samples.

Model-Based Clustering

Model-based clustering [22] assumes data are distributed according to a mixture of normal distributions and attempts to find a partition in samples by making use of a combination of the EM algorithm and the Bayesian Information Criterion (BIC). In a nutshell, for each value of k (= number of clusters) the partition which maximizes the classification likelihood is searched, and then the BIC is estimated for each of these models according to different cluster shapes (contour of the density of objects in a clusters) and volumes (amount of space occupied by the cluster in a *p*-dimensional space, where *p* is the number of variables). The larger the BIC, the stronger the evidence for the associated model.

Cluster Stability

To assess reliability of clustering results we considered two indices proposed in literature: the R index from McShane *et al.* [23] and the index from Smolkin *et al.* [24].

The R index substantially involves clustering of "perturbated" data sets according to some noise and comparison of results with the partition obtained on the original dataset. Thus, we added noise to the original data and re-clustered samples according to noisy expression data, and compared clustering results on the noisy datasets with the original ones. That is, considering *k* as the number of clusters, for every *i-th* cluster (with $i = 1,2,\ldots,k$) in the original solution, we consider all possible pairs of objects (i.e. samples) assigned to that specific cluster and evaluate if they cluster together also in the noisy dataset. Supposing that for the *i-th* cluster there are n_i samples, the number of pairs to be compared is $m_i = n_i(n_i - 1)/2$, so for each cluster we can compute a measure r_i which is a ratio between pairs of subjects of cluster *i* in the original dataset that cluster together also in the noisy dataset (c_i) and all possible pairs of subjects of cluster *i* (m_i), that is $r_i = c_i/m_i$. Such a cluster-wise measure can be extended to include all clusters, so we define the R index as

$$R = \frac{c_1 + c_2 + \ldots + c_k}{m_1 + m_2 + \ldots + m_k}$$

The higher the value of the index, the stronger is the "robustness" of clusters found in the original dataset. A relevant gain in reliability of this measure can be obtained by simulating several noisy datasets (say M) and computing the R index M times, averaging over the M values. To simulate the M noisy datasets in this work, we chose to add Gaussian noise (as suggested in [23]), by adding random values sampled from a Normal r.v. with zero mean and a specific standard deviation to the original expression values. McShane *et al.* [23] suggest to use the median of the gene-wise standard deviations, but they refer to a classical microarray experiment where only a few genes out of thousands are expected to show patterns of differential expression between experimental groups or clinical conditions. Since we expect a lot of genes to be differentially expressed between different types of tumors, we decided to use the 25th percentile of the distribution of gene-wise standard deviations, equal to 1.0696. Note that singleton clusters in the original solutions had to be treated differently: across all perturbed dataset we evaluated how many times each sample constituting a singleton clustered on its own also in the noisy datasets, and merged this information with those regarding pairs of sample from the other clusters. Chosen value of M was 1000 (number of perturbated datasets created).

The index from Smolkin *et al.* considers subspaces of the space of *p* variables and compares clustering solutions between the one obtained with *p* genes and the one obtained with a random subspace of $m < p$ genes. In particular, choose *m* genes randomly, perform clustering on this subset, evaluate whether clusters "re-appear" on these subsets (two additions or omissions are allowed), and perform these steps B times. The index, computed separately for each cluster, is the number of times that the cluster is found on all the B runs of the algorithm.

PRINCIPAL COMPONENTS ANALYSIS AND BIPLOTS

To evaluate more accurately the association between genes and samples Principal Components Analysis (PCA) was used as an exploratory multivariate technique which allows to reduce multivariate data to a lower dimensional space accounting for most of the variance of original data. This technique allows the visualization in a bi-dimensional space of the information on higher dimensional data, *via* the use of a PCA-based biplot.

In particular, given an *n x p* matrix X, the goal of PCA is to find m ≤ p uncorrelated linear combinations of the variables which explain most of the variation in X. These linear combinations will have the form

$$\alpha_k^{'} = \alpha_{k1}x_1 + \alpha_{k2}x_2 + ... + \alpha_{kp}x_p = \sum_{j=1}^{p} \alpha_{kj}x_j$$

where *k* indicates the general principal components and *j* the general variable. It can be shown that the α_k vectors of parameters, which we will refer to as loadings, correspond to the eigenvectors of ∑, the covariance matrix of X. The number of components that can be estimated is equal to the minimun between *n* and *p*, but in practice only those explaining the most variance will be considered: this translates in relevant reduction of the space of the variables.

To represent graphically results from this analysis, Gabriel suggested to use the biplot [9], a technique which allows to show variables and samples simultaneously on the same plot by means of a suitable rescaling. In particular, using Singular Value Decomposition (SVD) it is possible to write the X matrix as

$$X = USV^{'}$$

where, U (*n x r*) and V (*p x r*), are matrices whose columns form orthonormal basis , S is an *r x r* diagonal matrix whose elements $s_1^{1/2} \geq s_2^{1/2} \geq ... \geq s_r^{1/2}$ are the singular values of X, and *r* is the rank of X. If we define S^{α} for 0 ≤ α ≤ 1 as the diagonal matrix whose elements are $s_1^{\alpha/2} \geq s_2^{\alpha/2} \geq ... \geq s_r^{\alpha/2}$ and similarly for matrix $S^{1-\alpha}$ and let $G = US^{\alpha}$, $H^{'} = S^{1-\alpha}V^{'}$ then

$$GH^{'} = US^{\alpha}S^{1-\alpha}V^{'} = USV^{'} = X$$

So, the *(i,j)-th* element of X can be written as $x_{ij} = g_i^{'}h_j = \sum_{k=1}^{r} u_{ik}S_k^{1/2}v_{jk}$, which can be approximated by

$${}_m x_{ij} = \sum_{k=1}^{m} u_{ik}s_k^{1/2}v_{jk} = \sum_{k=1}^{m} g_{ik}h_{jk} = g_i^{*}h_j^{*}$$

where g_i^{*}, h_j^{*} contain first *m* elements of g_i and h_j respectively. This means that by plotting g_i^{*} and h_j^{*} on the same graph one can deduce several information about relationships between variables and subjects and among variables themselves [3].

A researcher could possibly be interested in understanding how new samples are projected on the biplot previously described, by projecting them on the PCA-biplot of X. Recalling that $X = USV^{'}$ and that both U and V have orthonormal columns, so that U'U = I and V'V = I, than it is possible to write

$$XV = USV^{'}V = US = G_X$$

and so the projection matrix G for a new dataset Y, whose columns contain informations on the same covariates as X, can be computed as $G_Y = YV^{'}$. The first two columns of G_Y are the projection coordinates of samples in Y on the PCA-based biplot of X. By plotting the new coordinates on the biplot it is possible to evaluate the association of the validation samples with the original ones.

There is, however, a problem which in literature is referred to as “batch effect”, that is, data from different studies can not be compared directly because of intrinsic differences due to different study setting (laboratory, tissue material, reagents, etc.) and to systematic bias not corrected by the normalization algorithm [25]. Since it is expected that the within-sample ordering of the *p* gene expression values will be similar for comparable samples also from two different studies, a ranking of these values within each subject may be adopted to compare samples across studies. Expression values are then replaced by their within-subject ranks for each subject for both the experimental and the validation dataset, and PCA is performed on ranked experimental data, visualizing results *via* a standard biplot. Validation is then performed graphically by passively projecting ranked validation data over this PCA-based biplot.

All analysis have been carried out using standard statistical software R 2.15.1 [26].

RESULTS

Experimental Data

Results from a gene expression study are presented. Expression values of 76 genes related to cell polarity and apical junctional complex components have been evaluated on 23 human tumor cell lines, of which 20 represent solid cancers arising from epithelial tissues (6 for breast, 8 for kidney and 6 for colon) and 3 derived from different kinds of leukemia~~s~~ (a systemic non-epithelial cancer), used as a negative control. Both genes and samples are a subset of a well-established publicly available dataset named NCI60, which contains information on thousands of genes of 60 human tumor cell lines [11,12]. As the process of loss of cell polarity is known to be connected to cancer development [17,18], goal of the study was to evaluate shared patterns of expression among different tumors all arising from epithelial tissues, and to explore associations between specific genes involved in such biological pathway and sets of similar samples. Characterization of samples is reported in Table **1**. Data were already normalized, and we only performed log2 transformation.

Table 1: Details of the 23 Cell Lines Involved in the Present Study

Names	Description
T47D	breast carcinoma
HS578T	breast carcinosarcoma
MDA-435	ductal breast carcinoma
MCF7	breast adenocarcinoma
BT-549	papillary infiltrating ductal carcinoma
MDA-231	breast adenocarcinoma
HCT-15	colon adenocarcinoma
HCC-2998	colon carcinoma
HCT-116	colon carcinoma poorly differentiated
SW-620	colon carcinoma
COLO205	colon adenocarcinoma
HT29	colon adenocarcinoma
RXF-393	kidney hypernephroma
A498	kidney adenocarcinoma
ACHN	renal cell carcinoma
CAKI-1	clear cell carcinoma
786-0	kidney adenocarcinoma
SN12C	renal cell carcinoma
UO-31	renal cell carcinoma
TK-10	renal spindle cell carcinoma
MOLT-4	lymphoblastic leukemia
CCRF-CEM	lymphoblastic leukemia
HL-60	promyelocytic leukemia

Cluster Analysis

Choice of k

In Figure **1** we show the plots for the choice of the number of clusters *k* using the different clustering algorithms. The silhouette plots of panel A (hierarchical) and B (k-means) suggest that a solution with 2 clusters is the best one (average silhouette of 0.484 and 0.526, respectively). Model-based clustering, on the other hand, provides a quite clear evidence that, irrespective of the supposed shape of the clusters, the solution with 4 groups outperforms the others with respect to the BIC (panel C). This is particularly true when clusters with equal volume and shape (orange dots) or with equal shape and different volume (brown dots) are considered (BIC = -5875.126 and -5851.330, respectively). Plot in panel D shows the number of clusters identified by AP algorithm in correspondence of a range of 1000 possible preference values ranging from -2.566 to 0.895. We can see that the more relevant vertical lines are associated with 2 and 3 clusters, with also a moderate evidence for a solution with 4 clusters.

Clustering Stability and Classification

The R index has been computed for all possible solutions from 2 to 7 clusters for all algorithms considered. Results are reported in Table **2**. Hierarchical solutions appear to be globally the more stable, across different values for k. Affinity Propagation and model-based appear to be the less stable solutions for all datasets considered, whereas k-means is comparable to hierarchical only for k = 2. Given the higher degree of stability, we decide to focus on hierarchical clustering solutions, and choose the solution for k = 2, reported in Table **3** and visualized in the heatmap of Figure **2**.

The index by Smolkin and Gosh [25] was computed for each of the 2 clusters defined in Table **3**, using a varying proportion of $m < p$ genes which ranged from 0.65 (49 genes) to 0.85 (65 genes). For each of these proportions we extracted 1000 reduced datasets,

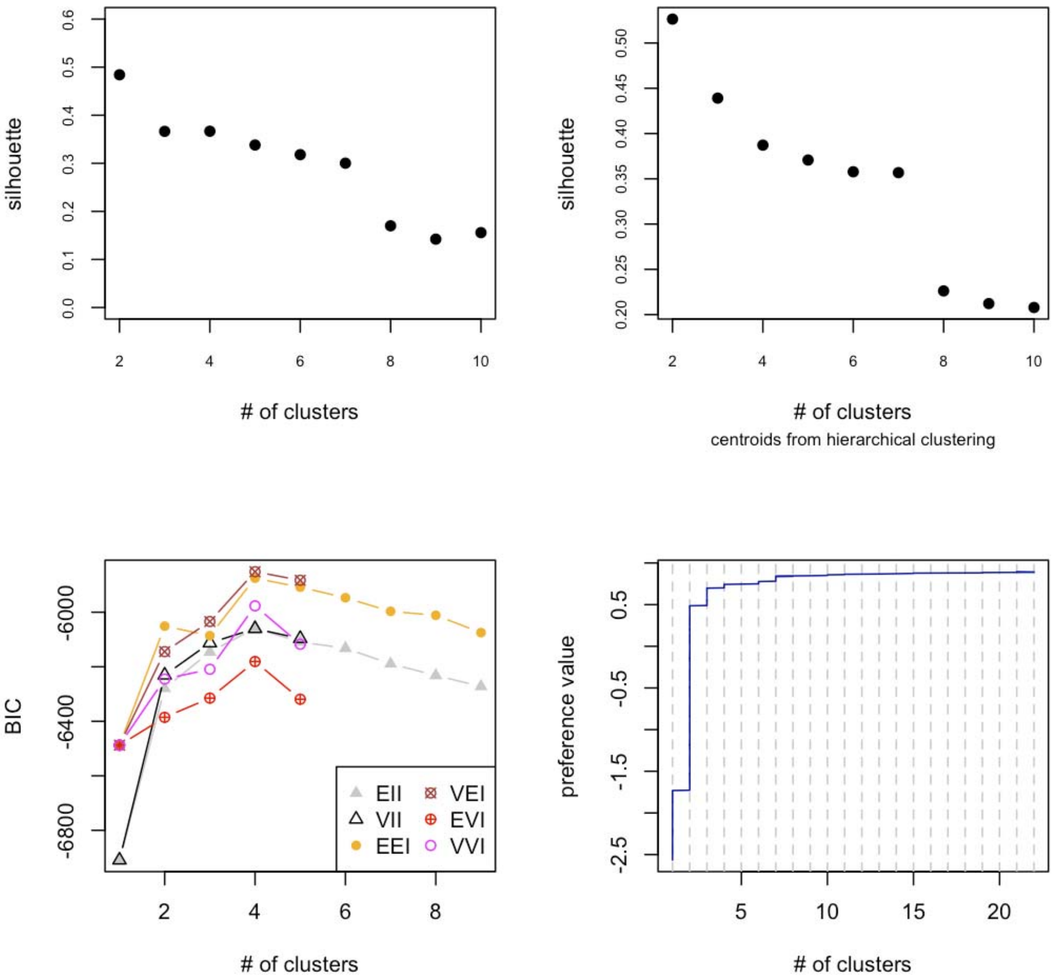

Figure 1: Choice of the number of clusters *k:* **a**) average silhouettes for hierarchical clustering; **b**) average silhouettes for k-means; **c**) Bayesian Information Criterion for model-based clustering; **d**) preference value plot for Affinity Propagation.

Table 2: R index for Cluster Stability [24], Computed on 1000 Perturbated datasets

k	Hierarchical	K-means	AP	Model-based
2	0.9923	0.9947	0.9426	0.8202
3	0.9509	0.8718	0.8087	0.812
4	0.9651	0.8094	0.7443	0.8142
5	0.9281	0.8527	0.6674	0.8197
6	0.9188	0.8526	0.601	0.7414
7	0.8499	0.803	0.6305	0.6768

Table 3: Classification of Cell Lines (k = 2) for Hierarchical Clustering

	Cluster 1	Cluster 2
Colon	6	0
Breast	2	4
Kidney	0	8
Leukemia	0	3

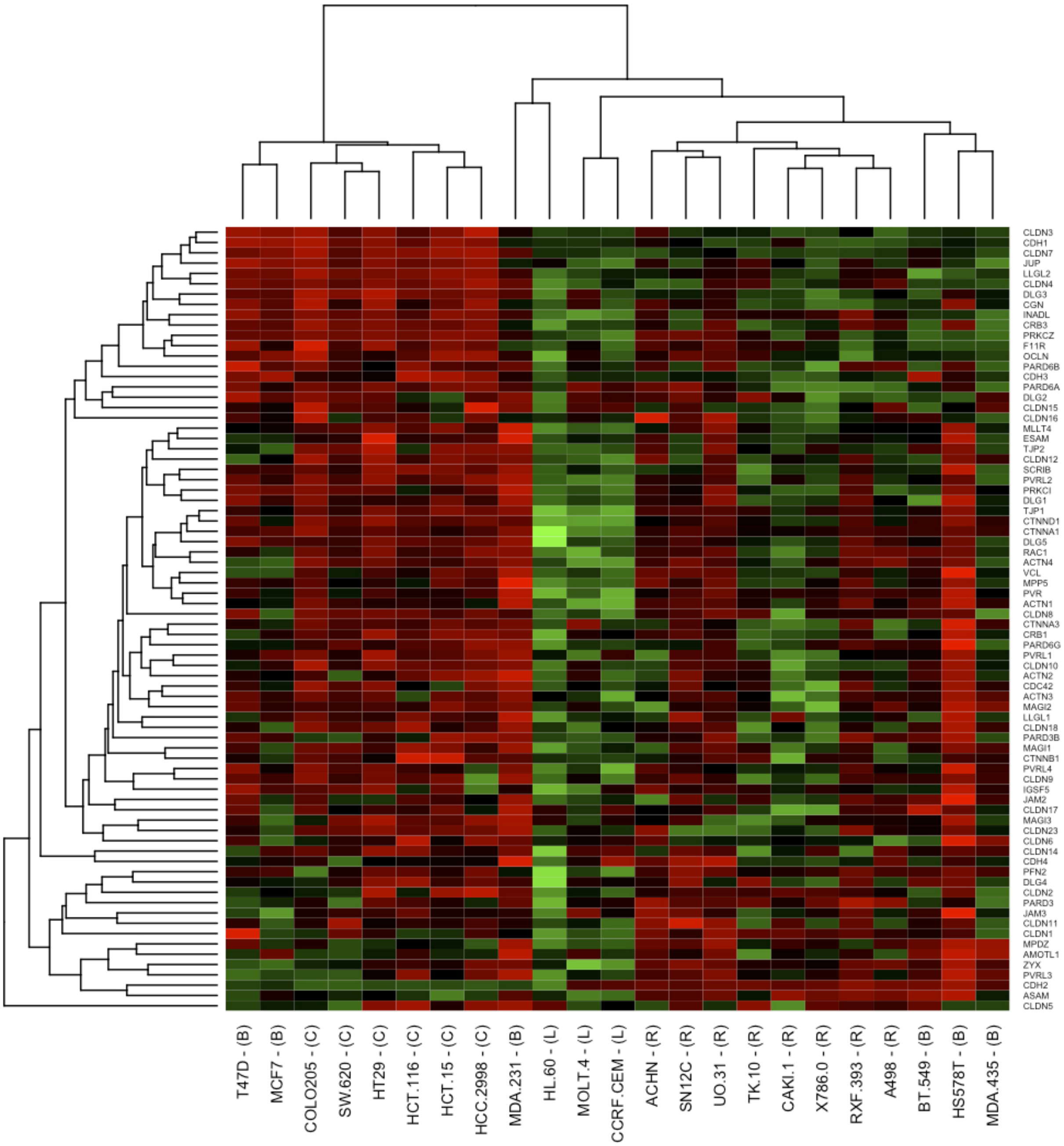

Figure 2: Heatmap of gene expression data from a subset of the NCI60 dataset. Row and column dendrograms were obtained using a hierarchical agglomerative algorithm, using average linkage and 1 - ρ^2 as distance metric. Green squares indicate low levels of expression, red squares high level of expression.

performed hierarchical cluster analysis on them using k = 2, computed the index for each cluster and then averaged over the datasets. Results are reported in Table **4**, and confirm the stability of this solution.

PCA-Based Biplots and Passive Projections

Principal Components Analysis (PCA) was performed on scaled and centered variables. Since we were not interested in doing some feature selection or in interpreting the PCs per se, we did not perform any further analysis about how many components to choose nor we did perform any rotation. The first two components accounted for 82% of total variability (67.8% and 14.2% respectively), meaning that a biplot, i.e. a graphic where information about subjects and variables is simultaneously plotted after some proper scaling (for further reference see [27]) is able to represent most of the variability of the data, thus giving very useful insight on the relationships between samples (or clusters of samples) and genes. The PCA-based biplot of our data is reported in Figure **3**.

Table 4: Results for the Index from Smolkin and Gosh [25]. Clusters Labels are those Specified in Table 3

m	Cluster 1	Cluster 2
49	0.899	0.899
50	0.931	0.931
51	0.919	0.919
52	0.943	0.943
53	0.947	0.947
54	0.960	0.960
55	0.955	0.955
56	0.972	0.972
57	0.973	0.973
58	0.979	0.979
59	0.989	0.989
60	0.984	0.984
61	0.987	0.987
62	0.990	0.990
63	0.995	0.995
64	0.998	0.998
65	0.997	0.997

It has to be pointed out that the samples in the upper-left portion of the plot are all those of cluster 1, colon lines and estrogen receptor-positive (ER+) breast lines, which show a positive association with a relevant number of genes, specifically those reported in Table **5**. Moreover, it can be noted that samples in cluster 2 are much more spread in the space of the first two principal components then those in cluster 1, which confirms that the genes chosen do not characterize them as much as for colon lines. These results seem to suggest a specific polarity profile for samples in Cluster 1, whereas no clear association can be retrieved for the remaining cluster.

Although cluster analysis suggests that cluster 1 is a strong structure and the biplot depicts a quite clear association structure between clusters and genes, it is relevant to know whether these results are reproducible or if they are some kind of technical artifact related to the NCI60 dataset considered here.

Three additional datasets, one containing information on 20 human renal cancer cell lines (including our 8 renal lines), and 51 genes, one including 3 lines derived from normal tissue (colon + kidney + breast) and 51 genes, and one including 3

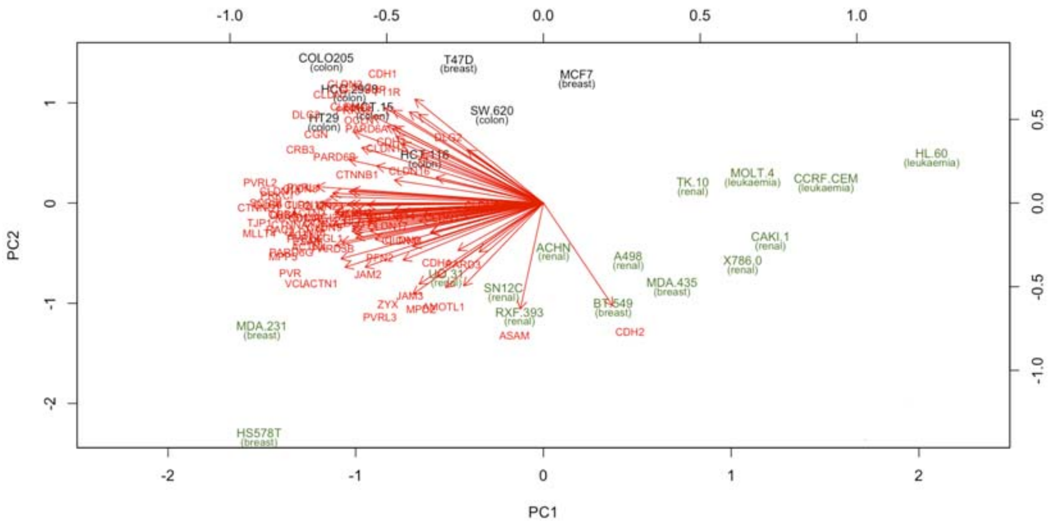

Figure 3: PCA-based biplot of NCI60 dataset. Black labelled samples belong to cluster 1, olive green labelled samples to cluster 2 (see Table **3**).

Table 5: Genes Associated to Cluster 1

CLDN15	PARD6A	JUP	PARD6B
OCLN	CLDN7	DLG2	CRB3
CLDN4	CLDN3	CDH1	
PRKCZ	LLGL2	INADL	
CDH3	F11R	CGN	

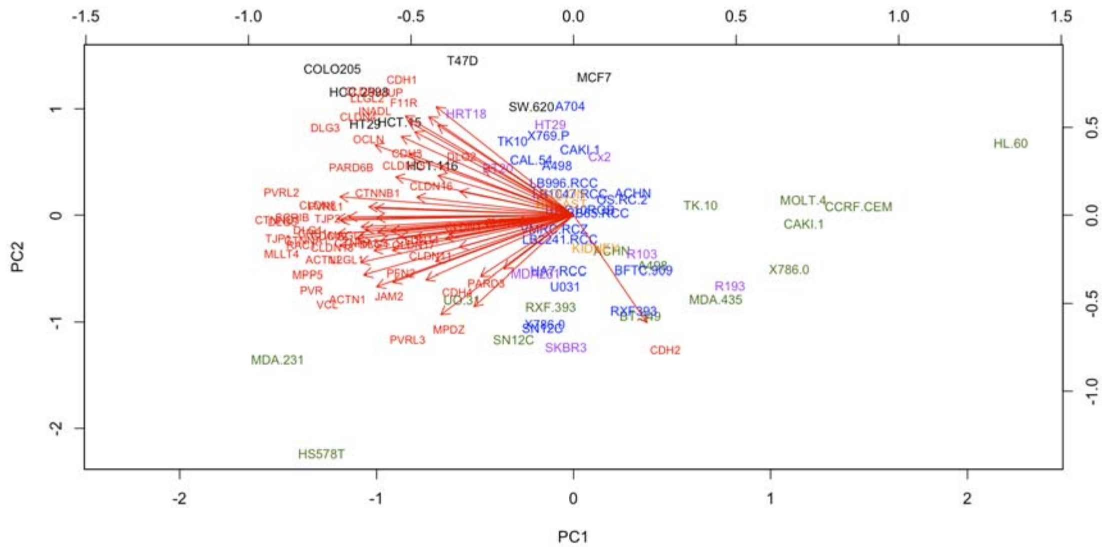

Figure 4: Passive projections of validation samples on the PCA-based biplot of NCI60 dataset. **Experimental samples:** Black labelled samples belong to cluster 1, olive green labelled samples to cluster 2 (see Table **3**). **Validation samples:** blue labelled samples come from the renal cancer validation dataset, violet labelled samples from the colon/breast cancer validation dataset and orange labelled samples from the normal tissues dataset.

colon cancer and 5 breast cancer cell lines with information on 62 genes were considered. Only the 51 genes common to all three datasets were used: this resulted in losing 25 genes, 33%, in the experimental dataset and 11 genes, 17.7%, in the colon-breast validation dataset. The validation samples were passively projected on the biplot of Figure **3**, obtaining the biplot in Figure **4**.

Clearly, renal cancer validation lines behave quite differently from experimental renal cancer lines. In general, there seem to be no variation along the first component for none of the validation datasets considered. Considering this plot, we could say that our experimental results substantially are not validated. However, the strange pattern, which is seen for validation lines in Figure **4**, is likely to be due to a batch effect, that is the association structure could be influenced by non-biological differences in the experimental settings. To properly validate results, we resorted to sample-wise ranked data (that is, gene expression values were ranked in ascending order within each sample for both experimental and validation datasets), visualizing biplot of PCA over ranked NCI60 data and then passively projecting the ranked validation data on this biplot. PCA over ranked

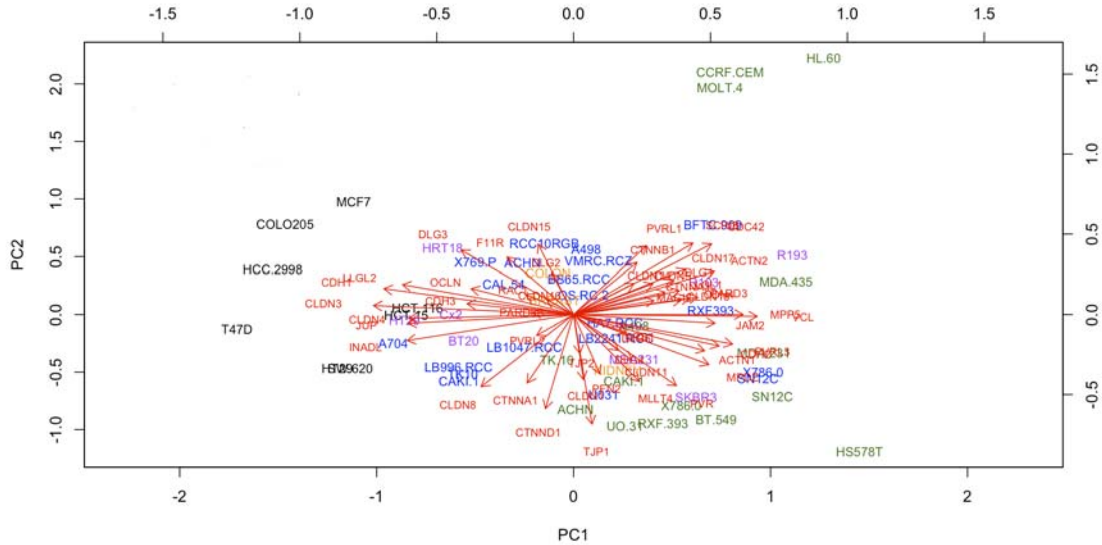

Figure 5: Passive projections of ranked validation samples on the PCA-based biplot of ranked NCI60 dataset. Sample labelling is the same as Figure **4**.

NCI60 data shows different features with respect to those on plain expression data, since here the first two components explain only 37% of the whole variability. In Figure **5** the “ranked” validation biplot is shown.

From this figure it can be seen that the structure of association depicted in Figure **3**, though being quite different, still conserves some specific patterns. In particular, lines in cluster 1 still group quite close and are associated with a set of 12 genes (namely INADL, JUP, CLDN4, CLDN3, CDH1, LLGL2, OCLN, RAC1, DLG3, PARD6B, CDLN16 and CDH3), ten of which were associated with this cluster also when using expression data (see Table **5**). It can also be noted that 4 of the colon/breast validation lines are close to the experimental samples of cluster 1; of these 4 samples, 3 are colon cancer lines and the fourth, BT20, is a breast cancer line which expresses ER mRNA despite its negative phenotype. The other breast cancer lines, all confirmed to be transcriptionally ER-negative, are on the opposite side of the plot, thus validating our experimental results. Validation renal cancer lines, on the other hand, show relevant patterns of dispersion in the space of first two principal components, but such a pattern is much more spread out than that seen in Figure **3**; moreover, of the 8 renal samples present both in the experimental and in the validation dataset, only 786.0, SN12C, UO31 and TK.10 show similar behaviour. Of the other 4 lines, some have completely different behaviour from validation to experimental (ACHN and A498), some show only moderate gaps.

DISCUSSION

The use of clustering methods in microarray research has so far become a standard in the analysis of biomolecular data, and many new clustering algorithms are being developed to deal with this kind of high-dimensional data. Connected with the choice of the clustering algorithm, the issues of choosing a proper number of clusters and of evaluating stability of the resulting classification are well known in microarray research. However, few papers deal with them, and rely on results from a single clustering algorithm, often choosing the number of clusters in an extremely subjective way and without any reliability assessment on clustering results. In this paper we have shown that using different clustering methods, with their related strategies to assess the number of clusters, is a useful way to choose a meaningful classification, and that evaluating clusters stability can provide deeper understanding of the robustness of the classification.

The issue of comparing results from different experiment is a very urgent topic in microarray research, where high-dimensional and noisy datasets pose a lot of critical issues that require proper statistical methods. To-date, some methods exist that deal with this problem: Parmigiani *et al.*, proposed a method named "integrative correlation", applicable to class comparison studies, i.e. studies when one is focusing on comparing expression profile between phenotypically different groups of subjects [28, 29]. Substantially, the proposal is to evaluate whether gene *j* is consistently expressed in different studies, one has to compute first the correlations of gene j with all other genes within each study, and then compute correlation between the within-study correlations. To determine whether such a measure reflects "reliability" of gene *j* across different studies, a null distribution is estimated and a cut-off is chosen according to the highest value of this null distribution. Lusa *et al.* [30] faced the issue of validating clustering results across different microarray experiments using various breast cancer dataset, claiming that “many difficulties remain in validating and extending class discovery results to new samples and that projection of clusters from one dataset to another must be done with care”.

In this paper we showed a method to validate results from a microarray experiment by making use of multivariate visualization methods related to Principal Components Analysis, and illustrated how this approach can be used to circumvent unwanted experimental effects which could confuse the relevant biological effects of interest by transforming data to their ranks. By applying the proposed technique to a subset of a microarray experiment on cancer cell lines we found that despite their common epithelial origin, colon, renal and breast cancer cell lines show a very different cell polarity profile. Specifically, we found a cluster of samples composed of colon cell lines, known to express estrogen receptors (ER+) [31], and of two ER+ breast lines. Notably, this cluster was found also by considering a number of clusters up to 7, for all algorithms considered, whereas the remaining lines showed different behaviour for larger values of *k* depending on the clustering algorithm. This cluster has a specific pattern of positive association with some epithelial markers such as CDH1, the gene encoding for E-cadherin [32, 33], and of negative association with CDH2, which encodes for N-cadherin, the typical marker associated with mesenchymal phenotype [34]. On the contrary, the cluster composed by renal cancer cell lines and ER- breast cancer lines, has an opposite pattern of association with epithelial and mesenchymal markers. This suggests that disruption of cell polarity and following alteration in the apical junctional complex may be influenced by other factors not strictly related to these processes, as can be seen by differential gene expression observed in ER+ versus ER- cell lines. Additionally, validation of results with passive projections shown in Figure **5** mostly confirmed results from previous analysis with regard to colon lines, whereas further investigations are needed for renal tissues, to better characterize their polarity profile and to evaluate influences of additional factors (e.g. hormone dependence).

With respect to cluster analysis, the main advantage, i.e. the comparison of results from different algorithms, could also be one possible limitation, in that the choice of the clustering algorithms to be compared is substantially subjective. It is possible that other techniques may give rise to different clusters with possibly different biological meaning, however since the cluster described above shows up throughout all algorithms for a large variety of *k* values we are quite confident in the biological relevance of our finding.

The main advantage of the proposed validation approach is that it provides useful and easy-to-interpret results in terms of association structure between samples and genes, yet there are some practical limitations. That is, the use of rank-based passive projections is likely to allow us to see only "the tip of the iceberg", since in the process of sample-wise ranking a lot of information contained in the expression data is going to be discarded, and thus only the robust association, both between samples and between samples and genes, will be confirmed and, eventually, validated. Additionally, increasing the dimensionality of data, in particular the number of genes involved, will make the plots overloaded and possibly uninterpretable. For such a reason, it is warranted that these methods are applied only when specific biological hypotheses have been formulated that allow the researcher to relevantly reduce the space of variables to be explored.

REFERENCES

[1] Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. New York: Springer 2003.

[2] Kaufman L, Rousseeuw PJ. Finding groups in data-An introduction to cluster analysis. New York: John Wiley and Sons, Inc 1990.

[3] Joliffe LT. Principal Components Analysis. 2nd ed. New York: Springer-Verlag 2002.

[4] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000; 97: 10101-6. http://dx.doi.org/10.1073/pnas.97.18.10101

[5] Chapman S, Schenk P, Kazan K, Manners J. Using biplots to interpret gene expression patterns in plants. Bioinformatics 2001; 18(1): 202-4. http://dx.doi.org/10.1093/bioinformatics/18.1.202

[6] Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data-analysis. Bioinformatics 2005; 21(15): 3201-12. http://dx.doi.org/10.1093/bioinformatics/bti517

[7] Datta S, Datta S. Comparison and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 2003; 19(4): 459-66. http://dx.doi.org/10.1093/bioinformatics/btg025

[8] Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. Bioinformatics 2001, 17(4): 309-18. http://dx.doi.org/10.1093/bioinformatics/17.4.309

[9] Gabriel KR. The biplot graphic display of matrices with application to principal components analysis. Biometrika 1971; 58(3): 453-67. http://dx.doi.org/10.1093/biomet/58.3.453

[10] Lander ES. Array of hope. Nat Genet 1999; 21: 3-4. http://dx.doi.org/10.1038/4427

[11] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, *et al.* Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000; 24: 227-35. http://dx.doi.org/10.1038/73432

[12] Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L *et al.* A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000; 24: 236-44. http://dx.doi.org/10.1038/73439

[13] Lee M, Vasioukhin V. Cell polarity and cancer-cell and tissue polarity as a non-canonical tumor suppressor. J Cell Sci 2008; 121: 1141-50. http://dx.doi.org/10.1242/jcs.016634

[14] Morrison SH, Kimble J. Asymmetric and symmetric stem-cell divisions in development and cancer. Nature 2006; 441: 1068-74. http://dx.doi.org/10.1038/nature04956

[15] Hugo H, Ackland ML, Blick T, *et al.* Epithelial-Mesenchymal and Mesenchymal-Epithelial Transitions in Carcinoma Progression. J Cell Physiol 2007; 213: 374-83. http://dx.doi.org/10.1002/jcp.21223

[16] Moreno-Buono G, Portillo F, Cano A. Transcriptional regulation of cell polarity in EMT and cancer. Oncogene 2008; 27: 6958-69. http://dx.doi.org/10.1038/onc.2008.346

[17] Cavallaro U, Cristofori G. Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. Nat Rev Cancer 2004; 4: 118-32. http://dx.doi.org/10.1038/nrc1276

[18] Cowin P, Rowlands TM, Hatsell SJ. Cadherins and catenins in breast cancer. Curr Opin Cell Biol 2005; 17: 499-508. http://dx.doi.org/10.1016/j.ceb.2005.08.014

[19] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998; 95(25): 14863-8. http://dx.doi.org/10.1073/pnas.95.25.14863

[20] Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007; 315: 972-6. http://dx.doi.org/10.1126/science.1136800

[21] Soria D, Garibaldi JM, Ambrogi F, Boracchi P, Raimondi E, Biganzoli E. Cancer profiles by Affinity Propagation. Int J Knowl Eng Soft Data Paradig 2009; 1(3): 195-215. http://dx.doi.org/10.1504/IJKESDP.2009.028814

[22] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 2002; 97(458): 611-31. http://dx.doi.org/10.1198/016214502760047131

[23] McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics 2002; 18(11): 1462-9. http://dx.doi.org/10.1093/bioinformatics/18.11.1462

[24] Smolkin M, Ghosh D. Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics 2003; 4: 36. http://dx.doi.org/10.1186/1471-2105-4-36

[25] Scherer A, Ed. Batch effects and noise in microarray experiments - Sources and Solutions. New York: Wiley 2009. http://dx.doi.org/10.1002/9780470685983

[26] R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[27] Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer-Verlag 2002. http://dx.doi.org/10.1007/978-0-387-21706-2

[28] Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E. A Cross-Study Comparison of Gene Expression Studies for the Molecular Classification of Lung Cancer. Clin Cancer Res 2004; 10: 2922-7. http://dx.doi.org/10.1158/1078-0432.CCR-03-0490

[29] Garrett-Mayer E, Parmigiani G, Zhong X, Cope L, Gabrielson E. Cross-Study validation and combined analysis of gene expression microarray data. Biostatistics 2008; 9(2): 333-54. http://dx.doi.org/10.1093/biostatistics/kxm033

[30] Lusa L, McShane LM, Reid JF, *et al.* Challenges in projecting clustering results across gene expression-profiling datasets. J Natl Canc Inst 2007; 99: 1715-23. http://dx.doi.org/10.1093/jnci/djm216

[31] Kennelly D, Kavanagh DO, Hogan AM, Winter DC. Oestrogen and the colon: potential mechanisms for cancer prevention. Lancet Oncol 2008; 9: 385-91. http://dx.doi.org/10.1016/S1470-2045(08)70100-1

[32] Heimann R, Lan F, McBride R, Heimann S. Separating favorable from unfavorable prognostic markers in breast cancer: the role of E-cadherin. Cancer Res 2000; 60: 298-304.

[33] Gould RBE, Bracken MB. E-cadherin immunohistochemical expression as a prognostic factor in infiltrating ductal carcinoma of the breast: a systematic review and meta-analysis. Breast Cancer Res Treat 2006; 100: 139-48. http://dx.doi.org/10.1007/s10549-006-9248-2

[34] Hazan RB, Phillips GR, Qiao RF, Norton L, Aaronson SA. Exogenous expression of NCadherinin breast cancer cells induces cell migration, invasion, and metastasis. J Cell Biol 2000; 148: 779-90. http://dx.doi.org/10.1083/jcb.148.4.779

Received on 25-09-2012 Accepted on 27-11-2012 Published on 21-12-2012

http://dx.doi.org/10.6000/1929-6029.2012.01.02.09