A Methodology for Analyzing SAGE Libraries for Cancer Profiling

JÖRG SANDER
University of Alberta
RAYMOND T. NG, MONICA C. SLEUMER, and MAN SAINT YUEN
University of British Columbia
and
STEVEN J. JONES
British Columbia Genome Sciences Centre

Serial Analysis of Gene Expression (SAGE) has proven to be an important alternative to microarray techniques for global profiling of mRNA populations. We have developed preprocessing methodologies to address problems in analyzing SAGE data due to noise caused by sequencing error, normalization methodologies to account for libraries sampled at different depths, and missing tag imputation methodologies to aid in the analysis of poorly sampled SAGE libraries. We have also used subspace selection using the Wilcoxon rank sum test to exclude tags that have similar expression levels regardless of source. Using these methodologies we have clustered, using the OPTICS algorithm, 88 SAGE libraries derived from cancerous and normal tissues as well as cell line material. Our results produced eight dense clusters representing ovarian cancer cell line, brain cancer cell line, brain cancer bulk tissue, prostate tissue, pancreatic cancer, breast cancer cell line, normal brain, and normal breast bulk tissue. The ovarian cancer and brain cancer cell lines clustered closely together, leading to a further investigation on possible associations between these two cancer types. We also investigated the utility of gene expression data in the classification between normal and cancerous tissues. Our results indicate that brain and breast cancer libraries have strong identities allowing robust discrimination from their normal counterparts. However, the SAGE expression data provide poor predictive accuracy in discriminating between prostate and ovarian cancers and their respective normal tissues.

Categories and Subject Descriptors: J3 [**Life and Medical Sciences**]—*Biology and genetics*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Gene expression, clustering, classification, cancer profiling

This research was supported by NCE IRIS and by NSERC Canada.
Authors' addresses: J. Sander, Department of Computing Science, University of Alberta, Edmonton AB, Canada T6G 2E8; email: joerg@cs.ualberta.ca; R. T. Ng, M. C. Sleumer, and M. S. Yuen, Department of Computer Science, University of British Columbia, Vancouver BC, Canada V6Y 1Z4; email: {mg, sleumer, myuen}@cs.ubc.ca; S. J. Jones, Genome Sciences Centre, British Columbia, Vancouver BC, Canada V5Z 4E6; email: sjones@bcgsc.ca.

1. INTRODUCTION

Gene expression profiling has been studied extensively in an attempt to determine the transcriptional changes, both causative and correlative, associated with the progression of cancer [Gray and Collins 2000]. Such an approach has the potential to determine new prognostic and diagnostic biomarkers in addition to new gene targets for therapeutic intervention.

1.1 Background

Cancers are typically classified based on macroscopic qualities such as the tissues they developed in. However, the same type of cancer can have different reactions to treatment in different people, and on the other hand, one treatment can be effective for cancers originating in different tissues. A possible explanation for these observations is that cancers, which are similar on a macroscopic level, may in fact be different at the sub-cellular level, and that certain cancers that originate in different tissues may actually be similar to each other at the sub-cellular level. The sub-cellular level of a tissue is characterized by very complex biochemical processes of which only a very small fraction is understood today. Most of the functions of a cell are performed by proteins that are produced by the cell via a mechanism called *gene expression*.

Gene expression is the process of making so-called messenger RNA (*mRNA*) copies of a gene (a part of the DNA strand of the cell). mRNA, which is essentially a sequence of four bases denoted by A, C, G, and T, is then translated in the cell into an amino acid sequence that forms the basis of a protein. Each tissue type in an organism requires different amounts of different proteins to perform its duties. The amount of a particular protein produced by a cell is partly controlled by the number of corresponding mRNA copies. The relative levels of mRNA of each gene in a tissue type are called the tissue's *gene expression profile*. This profile is assumed to be characteristic for a particular tissue type. Biochemists assume that many of the genes in the human genome are only expressed in one tissue type; but there are also so-called "housekeeping genes" that are expressed in all cells, for example, those that control transcription and translation.

Certain diseases, especially cancer, are caused by a sequence of errors that radically alter the normal pattern of gene expression. There may be a mutation in one or several genes that up- or down-regulate (i.e., control the rate of expression of) several other genes. For instance, the proteins that prevent uncontrolled cell growth and promote cell death are no longer produced, and the focus of the cell's gene expression machinery switches from producing the proteins appropriate to its tissue type to producing mostly proteins that are needed for cell growth and division. Theoretically, if the gene expression profile of a diseased tissue sample were known, it could be used to find out what has gone wrong in the cell and provide clues as to how to fix it.

There are two major methods of measurement for gene expression data, which produce a snapshot of the gene expression processes in a cell sample, at a certain point in time—the microarray method and the SAGE method.

In the microarray method, the sequences of the mRNAs that are measured must be known in advance. Many single-stranded pieces of DNA that

complement these short sequences are printed on a glass chip. The chip is then brought into contact with the mRNAs extracted from a cell sample. The mRNA in the sample binds to its DNA complement on the chip, and causes it to fluoresce. This fluorescence is detected with a laser. The more mRNA of a certain sequence is in the sample, the more its complementary spot on the chip will light up. An advantage of this method is that it is relatively inexpensive to produce large amounts of data. The major disadvantage, however, is that the experimenter must choose the mRNA sequences to be detected in a sample, and the sequences useful for cancer profiling may not be known.

Serial Analysis of Gene Expression (SAGE) [Velculescu et al. 1995] allows for the global profiling of an mRNA population, regardless of whether the transcripts have been previously identified. In the SAGE method, all the mRNA of a cell sample is collected, and a short subsequence, called a *tag*, is extracted from each mRNA (most commonly 10 base pairs in length, excluding the site for the restriction enzyme). These tags are then enumerated through standard DNA sequencing—the frequency of each tag is counted, giving the relative levels of the corresponding mRNA in the cell sample. The information about the frequencies of detected mRNA tags in a tissue sample is called a SAGE *library*.

The SAGE method is highly quantitative and SAGE profiles from different tissues can be readily compared. A substantive resource of SAGE data has been created as part of the SageMap initiative of the National Cancer Institute, USA [Lash et al. 2000], now part of the Gene Expression Omnibus [Edgar et al. 2002]. The availability of such a resource allows for a global comparison of normal tissues, cell lines and cancer types. Such a comparison may shed light on many interesting questions, such as:

- Is the SAGE method robust? Different libraries were created at different places by different groups. Is the SAGE method itself reproducible?
- Do different types of cancer behave differently at the gene expression level?
- Are there readily detectable subtypes of cancers?
- Is there such a notion that cancer type A is closer to type B than to type C?
- Can we identify, by computational analysis, a set of genes that characterize different types of cancer?

Frequently used methods to characterize and extract information from gene expression data are *clustering* (also known as "unsupervised classification/learning") and *classification* (also known as "supervised classification/learning" and in some information retrieval contexts as "categorization"). The goal of a clustering algorithm is to find unknown groups that may exist in a data set, whereas classification is supervised in the sense that classes are known beforehand and the methods try to learn a function that assigns objects correctly to one of the known classes (for an overview of different clustering and classification methods see, for instance, Han and Komber [2000].

There are many different types of clustering algorithms; the most common distinction is between partitioning and hierarchical algorithms. Partitioning algorithms construct a flat (single level) partition of a database D of n objects into a set of k clusters such that the objects in a cluster are more similar to each

other than to objects in different clusters. For a given number of clusters k, partitioning algorithms typically try to minimize an objective function such as the sum of the (squared) distances of objects to "their" cluster representative (i.e. the representative that is closest to them). Hierarchical clustering algorithms, on the other hand, do not actually partition a data set into clusters, but compute a hierarchical representation, which reflects its possibly hierarchical clustering structure. The well-known *single-link* method and its variants produce a so-called dendrogram, which is a tree that iteratively splits a data set into smaller subsets until each subset consists of only one object. Consequently, the root represents the whole data set, and the leaves represent individual objects. A different hierarchical clustering algorithm, which generalizes density-based clustering, is OPTICS [Ankerst et al. 1999]. This algorithm produces another type of output, a so-called reachability plot, which is a bar plot showing clusters as "dents" in the plot. In this article, we base much of our analysis on the result of this clustering method, which improves over traditional hierarchical clustering algorithms in several ways:

- The *single-link* or *nearest-neighbor* method, suffers from the so-called "single-link" or "chaining" effect by which clusters could be incorrectly merged if they are connected by a single line of points having the same inter-object distances as the points within the clusters (those "lines" often exist especially in data sets that contain background noise). OPTICS generalizes the nearest neighbor method in the sense that for a special parameter setting, the result of OPTICS is equivalent to the result of the nearest neighbor method. The same parameter, however, can be set to different values, which will weaken chaining effects and thus be able to better separate clusters in noisy data sets.
- OPTICS is more efficient on large data sets than traditional hierarchical clustering algorithms, if the dimension of the data set is not too high. For the 88 libraries we are working with in this article, however, this aspect is not important.
- The output of the OPTICS algorithm is a *reachability plot*, which in our opinion gives a clearer view of the hierarchical clustering structure and the density of the clusters than the typical dendrogram produced by traditional hierarchical clustering methods. However, a dendrogram can be generated from a reachability plot and vice versa [Sander et al. 2003].

The reachability plot is a simple bar plot, which visualizes a *cluster ordering* of the data set, where each library is represented by one bar. It shows simultaneously:

- which clusters are formed by which points (easily recognized as "valleys" in the plot),
- how dense clusters are in relation to each other (the deeper the "valley", the denser),
- a lower bound on the distance between 2 clusters (tallest bar between two "valleys"),

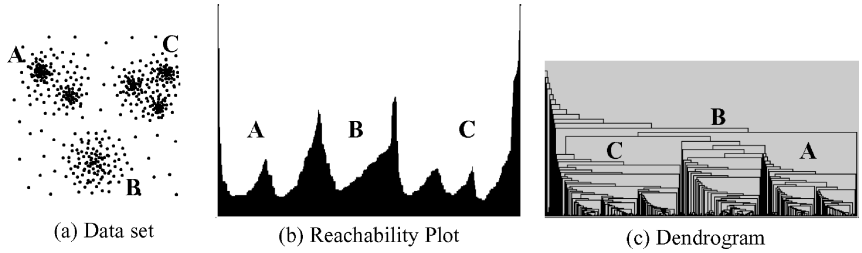


Fig. 1. Result of OPTICS for the given 2-dimensional data set (a), represented as a reachability plot (b), and a dendrogram (c). The regions in the plots corresponding to (nested) clusters are indicated by letters A, B, C.

- the hierarchical structure of the clusters ("nested valleys"), and
- which points are outliers (very large bars in the plot, not at the border of a cluster).

How to read dendrograms and a reachability plots is illustrated in Figure 1, using a simple 2-dimensional point data set.

1.2 Related Work

The SAGE method was introduced in 1995 [Velculescu et al. 1995]. The same group [Zhang et al. 1997] also proposed that the SAGE method could be used to study the differences between cancerous and normal cells. However, they only provided a brief example of the analysis that could potentially be done and did not provide the results of such an analysis.

In 1999 a website (http://www.ncbi.nlm.nih.gov/SAGE) was created as an offshoot of the Cancer Genome Anatomy Project (CGAP) at the National Center for Biotechnology Information (NCBI). CGAP is dedicated to collecting data on the genetics of cancer. An introduction to CGAP and an explanation of its purpose and the tools it contains is presented in Strausberg et al. [2000]. The purpose of the SAGE website is to provide data to the public, so that researchers could benefit from the SAGE technique without having to bear the expense of creating all of the data themselves [Lal et al. 1999]. Since then, various laboratories have submitted SAGE data to the site, both cancerous and noncancerous, from 10 different tissue types. The website also contains various tools to help researchers analyze the data [Lash et al. 2000]. However, it is hard to perform any analysis of the tag frequency distribution on a larger set of libraries.

The use of a clustering algorithm to study gene expression data was first proposed by the seminal paper of Eisen et al. [1998] in the context of microarray data. The authors applied hierarchical pairwise average-linkage clustering (using correlation coefficient as similarity measure) to the genes on microarrays for the budding yeast genome and a human fibroblast cell line. The same clustering methodology (a software is publicly available from the Eisen lab at http://rana.lbl.gov/) has since then been used for different analyses of microarrays, including the clustering of genes on microarrays for normal and cancerous breast tissue [Perou et al. 1999], the clustering of different tissues on microarrays for different types of lymphoma, in order to detect possible subtypes of

lymphoma [Alizadeh et al. 2000], and the clustering of both the tissues and the genes on microarrays for normal and cancerous colon tissue [Alon et al. 1999]. Other clustering paradigms have also been tried on microarray data, such as k-means (e.g., Tavazoie et al. [1999]), self-organizing maps (e.g., Golub et al. [1999]), model-based clustering (e.g., Yeung et al. [2001]), and specially designed algorithms such as the method of Ben-Dor et al. [1999] which is based on a biological model of genes.

Reports on cluster analysis of SAGE data have only recently started to appear (including some preliminary results, which we extend in this article [Ng et al. 2001]. Most approaches simply apply the software package from the Eisen lab. Porter et al. [2001] clustered both genes and libraries of 8 normal and cancerous breast tissues, finding differences in the variability of gene expressions in normal and cancerous cells. Nacht et al. [2001], clustered 9 normal and cancerous lung tissues, and showed that normal and cancer could be separated based even on only the 115 most abundant tags. Van Ruissen et al. [2002] used two-way clustering of both libraries and genes of different skin tissues, finding two clusters of genes that are up-regulated in cancerous tissue. Hashimoto et al. [2003] applied hierarchical clustering to genes of different types of leukocytes, showing that genes are differentially expressed in each leukocyte population, depending on their differentiation stages. Buckhaults et al. [2003] studied 62 surgically removed samples of 4 different cancers (primary cancers and secondary metastases of ovary, breast, pancreas, and colon) in order to predict the origin of a cancer. They first selected only the top-ranked tags identified by a support vector machine classifier: tags that best separated between the four classes. Based on these tags that best distinguish between the four classes they applied two-way clustering to both genes and tissues detecting that metastases of a cancer clustered together with their corresponding primary tumors.

1.3 Overview of Our Contributions

The previous clustering studies of SAGE were in general performed by different labs that produced their own (typically very small sets of) libraries. The main contribution of this article is a methodology for analyzing heterogeneous SAGE libraries on a large scale.[1] We first develop a sequence of steps to cleanse or preprocess the libraries including missing tag imputation and subspace selection (i.e., removing tags that may just be noise, and selecting tags that may be more discriminating). We show that all four steps have to be applied together for an effective subsequent analysis of the data—none of the steps alone can produce satisfactory results.

After the libraries have been cleansed, we perform hierarchical clustering using OPTICS on the libraries. Apart from clustering, we also apply nearest-neighbour classification. This is to examine whether a specific type of cancer has a strong identity or "signature" at the gene expression level.

We show that our methodology is effective in that it sheds light on some of the biological questions we explore. Our analysis suggests that the SAGE technique is robust. For instance, brain cancer libraries developed in different

[1]A preliminary version of this work appeared as Ng et al. [2001]

laboratories form a strong cluster, suggesting that results obtained in different laboratories are in fact comparable. Our analysis also indicates that brain and breast cancer have much stronger gene expression signatures than ovarian and prostate cancer do. Finally, our analysis finds one strong cluster of brain and ovarian cancer libraries, suggesting that there may be interesting similarities between the two cancer types. We also identify a set of genes that distinguish this cluster from normal brain and ovarian tissues. In many cases, literature searches confirm that the identified genes are promising candidates for further analysis.

2. RESULTS

The following analysis is based on 88 SAGE libraries, which are publicly available on the NCBI SAGE website as of January 2001. Each of these SAGE libraries is made of a sample from one of the following tissues: brain, breast, prostate, ovary, colon, pancreas, vascular tissue, skin, or kidney. The other information consistently included with each library is whether it was made from cancerous or normal tissue, and whether it was made from bulk tissue *in vivo* or a cell line grown *in vitro*. The data in each library includes a list of 10-base tags that were detected in the sample, and the number of times each tag was detected. The number of *unique tags* in a library is the number of different tags that were detected in the sample, each of which was detected with some integer frequency. The *total number of tag copies* is the sum of all the frequencies of all the tags in a library.

In analyzing the SAGE data, four major features need to be accounted for. The first three issues are a consequence of the sampling and sequencing errors associated with the SAGE method [Stollberg et al. 2000]: First, the SAGE method is highly prone to sequencing error, which creates a large amount of noise and obscures the clustering structure. Second, the libraries differ largely in terms of total number of tag copies due to differences in the depth to which individual SAGE libraries were sampled. Third, some of the SAGE libraries have been subjected to very limited sequencing, having 0 values for most of the tags. The fourth issue is that the "raw" data is extremely high dimensional. For our cluster analysis, each tag corresponds to a dimension. Thus, if we consider 50,000 tags simultaneously, each library corresponds to one point in the 50,000-dimensional space. But it is not clear which of the 50,000 dimensions are relevant for cancer profiling. To deal with these problems, we designed four preprocessing steps: error removal, normalization, missing tag imputation, and subspace selection. Each of these steps is discussed in detail below.

2.1 Removing Erroneous Tags

If we assume that the single pass sequences used to generate the SAGE tags have a base error rate of approximately 1% per base, then we can expect that approximately 10% $(1.00\text{-}(0.99)^{10})$ of the total number of tag copies in each SAGE library will contain at least one sequencing error (an error can occur at each of the 10 base positions). Since these errors result in noise and increase the dimensionality of the data, our first preprocessing step is error removal.

Within one library, up to 80% of the unique tags have a frequency of 1, but these make up about 20% of the total number of copies of tags. Some of the single frequency tags represent genuine low frequency mRNAs. However, it is impossible to tell which of the single frequency tags are errors by only looking at a single library.

Since the size of the set of possible tags is $4^{10} \approx 10^6$, and only about 3×10^5 of these tags have ever appeared in any library, a sequence error in a tag most likely represents a completely new tag that never appears in any other library. Therefore, our method of removing erroneous tags without removing legitimate single-frequency tags follows a thresholding approach (as used, e.g., by Hashimoto et al. [2003]): remove those tags that have a frequency of no more than 1 in all the libraries. In this manner, tags that have a frequency of 1 in some libraries but a higher frequency in other libraries are not removed. This approach taken to removing erroneous tags is similar to that now done by the SAGE genie software, which generates a "confident tag list" (see Boon et al. [2002]).

The total number of unique tags across all libraries (which is the dimensionality of the subsequent space upon which the libraries are clustered) before tag removal is about 350,000; after the removal it is 58,524, which is a significant reduction of the dimensionality of the data set as well as the noise. In each library, between 5% and 15% of the total number of tag copies are removed, which is in the range of the expected number of sequencing errors per library.

2.2 Normalization

The 88 SAGE libraries upon which we base our analysis were made by different institutions. Because the amount of resources spent by these institutions for sequencing varies widely, each library has a different total number of tag copies. As a consequence, it is not meaningful to use some of the distance functions such as the Euclidean distance to compare the libraries to each other, since their tag frequencies are not on the same scale. Thus, we normalized the tag frequencies by simply scaling them all up to the same total number of tag copies (e.g., tags per million), which is a common way to compensate for unequal tag numbers in SAGE data analysis (see, e.g., Porter et al. [2001]). For other distance functions such as the correlation coefficient, this type of scaling does not have any effect.

2.3 Preliminary Attempts

In our early attempts to cluster the SAGE libraries, we investigated the use of CLARANS [Ng and Han 1994], a k-medoids algorithm, which is a partitioning algorithm that tries to find k objects as representatives of clusters so that the sum of all distances of all objects to their closest representative is minimized. We focused on a single tissue type: 14 breast tissue libraries that were available at the time. We applied some simple screening procedures to reduce the dimensionality of the data set, such as requiring a minimum frequency threshold for tags. We tested several similarity and dissimilarity measures such as Euclidean distance, and the correlation coefficient. This analysis showed that the breast tissue libraries form clusters, but these clusters were formed according

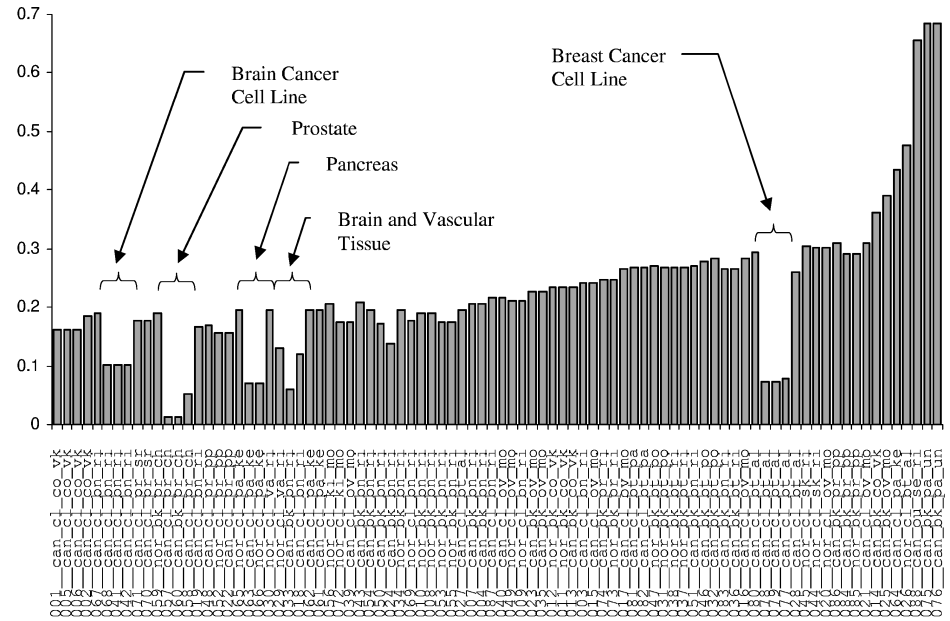
to tissue source (cell line or bulk tissue) and not according to neoplastic state (cancer or normal). All of the bulk tissue-derived libraries went into one cluster while all the cell line-derived libraries went into another cluster. But even after eliminating all tags that were only found in bulk tissues, the success in separating cancerous breast tissue from normal breast tissue was only moderate. Furthermore, CLARANS requires the user to specify the number of clusters to be formed. From a biological standpoint, this number is hard to determine *a priori*.

Because of these weak results, and because we actually wanted to see whether some types of cancers were related to each other at the gene expression level as well as to find clusters within one category, we decided to use a hierarchical clustering algorithm for a more detailed analysis. In contrast to partitioning algorithms, hierarchical algorithms do not require the number of clusters to be specified *a priori*. Hierarchical algorithms were used in many previously published papers on the clustering of microarray data. For reasons discussed in Section 1.1, our analysis here is based on the hierarchical clustering algorithm OPTICS [Ankerst et al. 1999], and we will show both the reachability plots and (for readers more familiar with the interpretation of tree representations) the dendrograms of our clustering results.

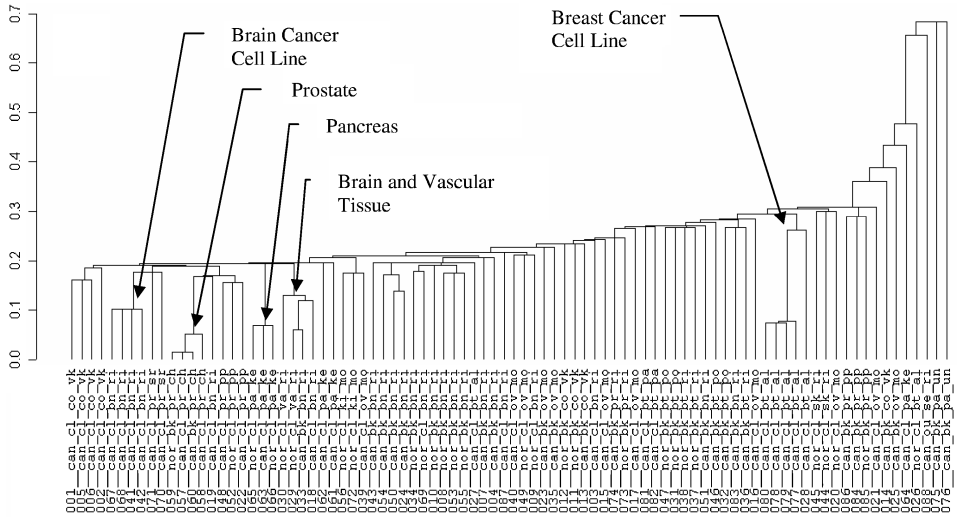
Applying OPTICS naively to the SAGE data (after error removal and normalization) does not produce a strong clustering structure. Figure 2a shows the reachability plot using correlation coefficient as a similarity measure. Figure 2b is the corresponding dendrogram. Each library is labeled by a library number (001 to 088), cancerous or normal indicator (*can*, *nor*), bulk tissue or cell line indicator (*bk*, *cl*), tissue type (*bn* = brain, *bt* = breast, *co* = colon, *ki* = kidney, *ov* = ovary, *pa* = pancreas, *pr* = prostate, *sk* = skin, *va* = vascular) and a code for the laboratory that created the library. For instance, the label 032-can-bk-bt-po corresponds to the cancerous, bulk tissue breast library whose number was 32 and was created by the laboratory labeled *po*.

There are very few clusters: a brain cancer cell line cluster, a prostate cluster, a pancreas cluster, a mixed cluster containing brain and vascular tissue, and a breast cancer cell line cluster. Each of these clusters contains only 3 to 5 libraries, all of which are of a single type of tissue (with the exception of the cluster containing both brain and vascular tissue). The other 56 libraries do not group into pronounced clusters.

Note that there are several different possibilities for a similarity or dissimilarity measure between SAGE libraries, such as the Manhattan distance, the Euclidean distance, and the correlation coefficient. The one that is most commonly used in the literature for microarray data is the correlation coefficient (c.f. Section 1.2). We clustered the SAGE libraries using all of the mentioned functions and found that the clustering structures are typically very robust with respect to the similarity function—they differ in general only slightly. However, the correlation coefficient always resulted in the most pronounced structure, and is also consistent with previously published work. Therefore, we will present our results using the correlation coefficient. Furthermore, note that we do not use the log scale as in typical microarray analyses. Intensity values on a microarray chip represent the ratios of gene expressions of



(a) Reachability Plot



(b) Dedrogram

Fig. 2. Result of OPTICS after error removal and normalization.

a condition relative to a reference condition. These values are typically log transformed, mainly so that equivalent fold changes in either direction (over- or under-expression) have the same absolute value (x/y and y/x are very different, whereas $\log(x/y) = -\log(y/x)$). In SAGE, tag counts represent absolute number of occurrences of a gene tag, so a log transformation is not necessary.

The weak clustering result is consistent with our preliminary results using CLARANS and indicates that we have to address the problem of missing values due to incomplete sequencing and suitable subspace selection.

2.4 Missing Tag Imputation

After the outlined error removal, the total number of tag copies in a library varies from 1293 to 79498. Consider all the libraries whose total tag counts are below 30,000. Out of the 88, there are 28 such libraries. If we were to discard these 28 libraries, we would have lost valuable data contained in them. If we were to normalize the libraries just based on the existing tag counts, the counts of some of the tags would have been exaggerated. The strategy we adopt is to "impute" the counts of the missing tags. Specifically, for a given library with a low total count, we use similar libraries to conservatively estimate the count of a missing tag. This estimate allows us to include the library and its true tag counts in our analysis. While the details of our imputation strategy are given below, we basically use a method similar to those used for filling in answers on incomplete surveys (see http://www.utexas.edu/cc/faqs/stat/general/gen25.html).

Note that missing value imputation is not new to gene expression analysis. But the estimation methods for DNA microarrays as proposed in Troyanskaya et al. [2001] are not applicable in this case, since missing values in DNA microarrays are due to different types of errors. Those methods look at the gene expressions of other genes in the experiments and use their values to estimate a missing value. For instance, the best method reported in Troyanskaya et al. [2001], the KNNimpute method, first selects the k genes that have the most similar expression profiles to the gene having a missing value under some condition; then a weighted average of the k most similar genes that have a value under the same condition is used as an estimate for the missing value. This method is based on the assumptions that errors occur randomly and locally at specific spots on the microarray, and that the gene with the missing value is over- or under-expressed by a similar factor as a small cluster of genes. Both assumptions are not applicable in the case of our SAGE data set. Here, missing values are systematic in the sense that an incomplete sequencing affects all genes in a specific condition or experiment. Furthermore, the conditions or experiments involve very different tissue types where we cannot assume that every gene is co-expressed in the same cluster of other genes in all tissues, or that the absolute mRNA counts of co-expressed genes are similar (even though they may be up- or down-regulated by the same factor).

Therefore, in our approach, we adjust the frequency of tags in a library that has a low total number of tag copies by using more complete libraries in the same category—from the same tissue, source, and neoplastic state. For our study, libraries containing 30,000 or fewer tags underwent this imputation, the underlying assumption of this process being that it is more conservative to consider a tag to be absent due to insufficient sampling alone in cases where the

same tag has been observed in libraries of similar origin sampled at a greater depth. We intend to adjust the truncated SAGE libraries as conservatively as possible in order not to introduce artificial similarities between them. Our method for replacing missing tags is as following:

- Identify libraries with a low total number of tag copies. The threshold value was set to 30,000 in consultation with experts familiar with the SAGE method, subject to the constraint that not more than 30% of the libraries were actually affected.
- For each of these,
  - Create a list of libraries in the same category, which have a total number of tag copies exceeding 30,000.
  - Scale them down so that the total number of tag copies matches the total number of tag copies of the low tag library. This gives an estimate of the expression values in those libraries as if they were sequenced incompletely to the same degree as the truncated libraries of interest.
  - Replace each frequency 0 in the low-tag library with the lowest non-zero frequency of that tag in the scaled down libraries. Since the intention here is to solve the problem of the truncated libraries as conservatively as possible, we replace the zeros with the lowest possible value from the scaled down libraries, instead of, for instance, with a weighted average, or by replacing the values in a truncated library with the values of the most similar down-scaled library in the same category.

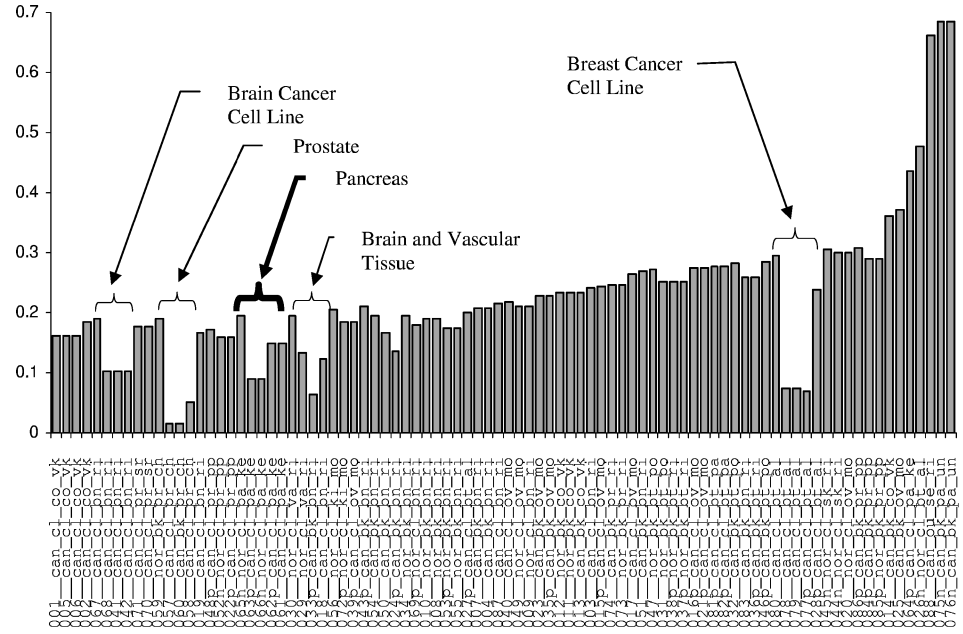
Of the 88 SAGE libraries studied, 28 were subjected to tag imputation. After the low frequency tags were removed, the missing values were filled in, and the libraries were normalized, the resulting data was clustered by OPTICS. Figure 3a (reachability plot) and 3b (dendrogram) show the result, again with correlation coefficient as similarity measure.

Basically, the same clustering structure as before is found; only the pancreas cluster now includes 2 additional libraries. The imputation of missing values did not change the data radically with respect to the clustering structure, and did not create any "artificial" clusters by making libraries too similar to each other. However, the imputation of missing values has important benefits for subspace selection.

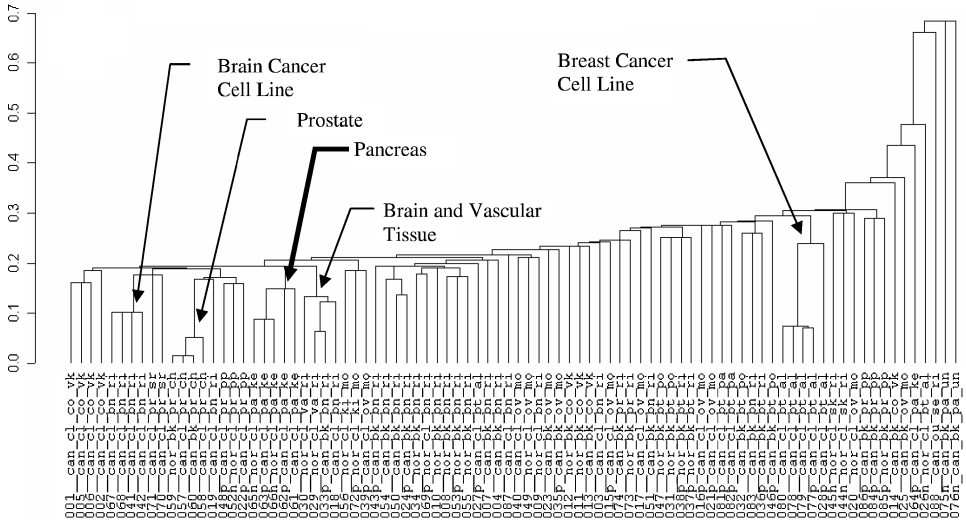
2.5 Subspace Selection

So far, we have used the whole set of tags when computing the similarities between libraries on which a clustering algorithm is based, and the resulting clustering structures using this whole space of features (tags) were very weak. Thus, we embark on subspace selection, the goal of which is to identify a subset of tags that discriminate between the following two kinds of tags:

- Tags that have similar expression levels in all or most of the libraries, independent of neoplastic state, source and tissue type.
- Tags that have different expression patterns in different situations.



(a) Reachability Plot



(b) Dedrogram

Fig. 3. Result of OPTICS after error removal, normalization, and missing tag imputation.

The first kind of tags does not help in forming clustering structures. It even produces a dilution effect to the true clusters, by disturbing the similarity and dissimilarity of libraries. Thus, these tags are removed and only the discriminating tags are kept for further analysis.

The key question here is how to select discriminating tags. Since the total number of tag copies in our data set is extremely high, methods that search over subsets of tags are computationally too expensive, and we have to restrict ourselves to a method that checks properties of single tags. A similar strategy has also been employed in the context of cancer classification using DNA microarray data sets. These methods typically compute a value for each gene that measures the gene's suitability for class separation. The genes are ranked according to this measure, and the top m, or all genes with a value above a certain threshold t, are selected. The values for m or t have to be specified by the user. Definitions that have been proposed for measuring a gene's relative class separation include, for instance, Golub et al. [1999] correlation metric based on the means and the standard deviation of the expression values for a gene in the different given classes, and Ben-Dor et al.'s [2000] TnoM score, which measures the, classification error when assigning the samples to the classes only by comparing the expression value for the gene of interest to a learned threshold. Those methods have some drawbacks in our context because our data set contains samples from very different tissues and the correlation of any tag with the classes cancer or normal might be very low, which in turn would make it very difficult to specify the required threshold parameters for the methods.

We addressed the issue of subspace selection by using the Wilcoxon rank sum test [Wilcoxon 1945], which can tests whether two samples are taken from the same population. In addition to not requiring input parameters, it is a nonparametric statistical test—it makes no assumption concerning the distribution of the two classes, and it can be successfully applied even to very small samples. For all tests a significance level of α equal to 0.01 was used.

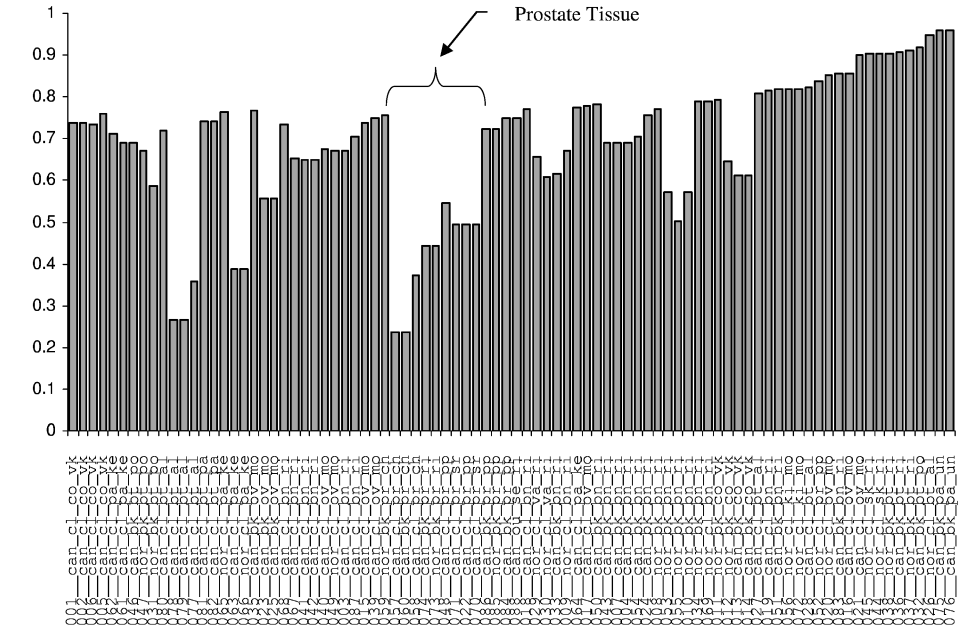
In our application, we want to detect possible similarities between cancers in different tissue type as well as possible subtypes of a cancer in a particular tissue type. Therefore, in order to avoid biasing the result, we cannot use tissue specific differences in expression to reduce the number of tags. Therefore, to select a relevant subspace of tags, we first applied the test on a tag by tag basis to determine those tags that have significantly different expression levels in cancerous versus normal cells, without regard to the tissue type. For the Wilcoxon test, we look at two groups of tissues at a time (e.g., cancer versus normal), and we consider the frequencies of a tag T in the libraries of each group as a sample. We then apply the test to these two samples to determine whether the (null) hypothesis that the two samples are taken from the same population can be rejected (at the given confidence level). We retain a tag only if its expression in cancerous and normal cells is significantly different. The intuition behind this heuristic selection of a subspace of tags is that we only want to select "informative" tags and remove those that do not have any individual predictive ability (according to our statistical test). In particular, if the distribution of the expression values of a tag are so similar in both classes (e.g. cancer vs. normal) that we cannot reject the hypothesis that the values follow the same distribution (according to the Wilcoxon test), their (individual) ability to distinguish between the two classes is consequently very low and we do not include this tag in the final subspace.

As described earlier, our preliminary results indicated that the tissue source (i.e., bulk tissue, or cell line) can dominate the similarity between the tissues. We tried to weaken this kind of distortion by applying the test again to also determine the tags that have significantly different expression levels in bulk tissue versus cell line. Those tags were then subtracted from the set of potentially cancer relevant tags selected in the first step.

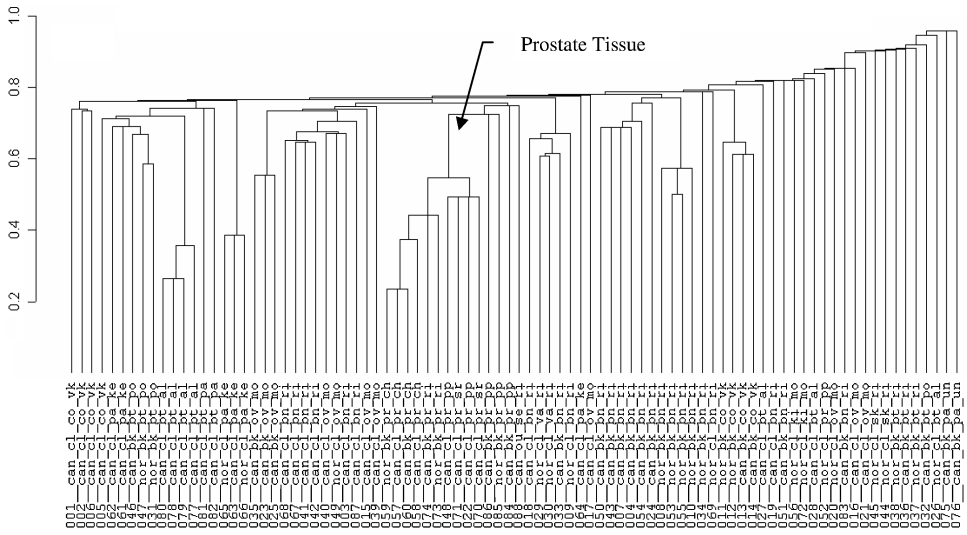
To demonstrate the effect of our missing value imputation on the subspace selection method, we first applied the Wilcoxon test to the SAGE data after error removal and normalization, but without imputing missing values. The Wilcoxon test selected 45,354 tags out of 58,524 when testing cancer versus normal. 722 tags out of the original 58,524, were selected when testing bulk tissue versus cell line. Only 49 of these tags were also present in the 45,354 tags selected by the first test. We removed these tags from our dataset, leaving us with a dimension of 45,305. When clustered with OPTICS, we obtain the result shown in Figures 4a and 4b. The subspace selection resulted in a slightly more pronounced clustering structure than in the previous experiments. For instance, the figure shows a large prostate tissue cluster containing 13 libraries. However, most of the clusters are still very small and several of them are mixed with respect to tissue type and neoplastic state. Furthermore, a large number of libraries are not contained in any significant cluster.

Finally, the result of OPTICS on the SAGE data using *all* the pre-processing steps in combination is shown in Figures 5a and 5b. In this case, the Wilcoxon test selected 40,016 tags when testing cancer versus normal, and 186 of these were also selected when testing bulk tissue versus cell line (i.e. 39,830 tags remained). In this result the clusters are more prominent and all of them except one are pure in terms of tissue type and neoplastic state.

- Eight dense clusters formed: ovarian cancer cell line, brain cancer cell line, brain cancer bulk tissue, prostate tissue, pancreatic cancer, breast cancer cell line, normal brain, and normal breast bulk tissue.
- For the soundness of our cluster analysis method for heterogeneous SAGE libraries, it is important to validate that the libraries do not form clusters by laboratories alone. Spurious clusters may be formed due to artificial laboratory effects, and the issue is complicated by the fact that many laboratories produced only libraries of a single tissue type. On this regard, it is reassuring that meaningful clusters consisting of libraries from different laboratories have been found, and that libraries produced by the same laboratory have been grouped into different clusters based on biological similarity.
- The prostate tissue cluster is almost pure—only 3 out of 12 libraries are normal while the rest are prostate cancer. However, since all of them were made from bulk tissue samples, it is still possible that these 3 normal libraries contain a lot of heterogeneity, presumably reflecting the difficulty in the clinical dissection of prostate tumors. The prostate cluster is also interesting in that the libraries came from different laboratories (i.e., pp, sr, ri and ch), strongly suggesting that this cluster is formed for biological reasons.
- The ovarian cancer cluster did not appear in the any of the previous results, and is surprisingly close to the brain cancer cell line cluster, which suggests

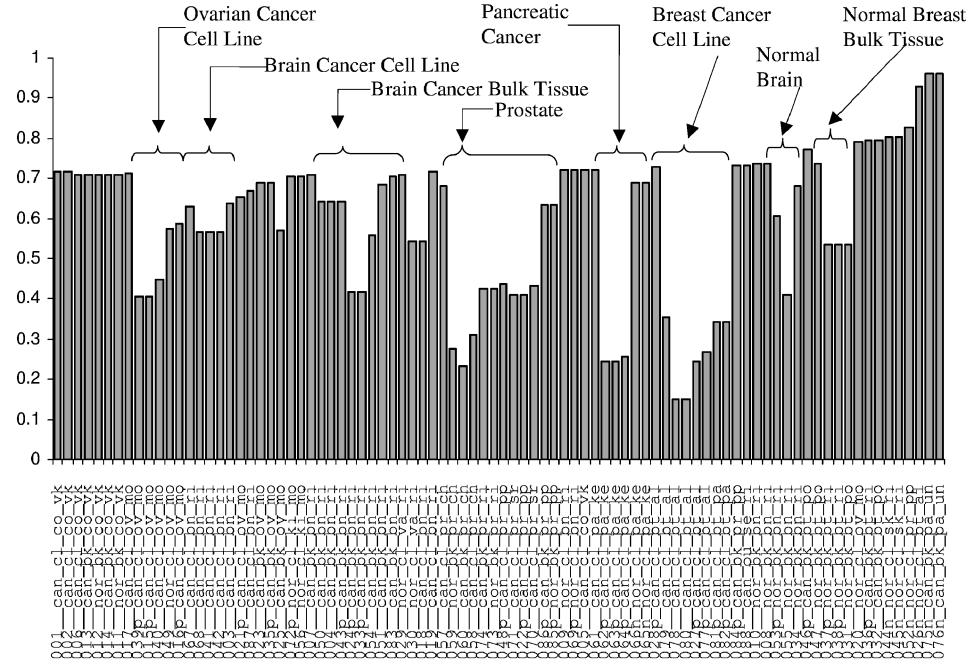


(a) Reachability Plot

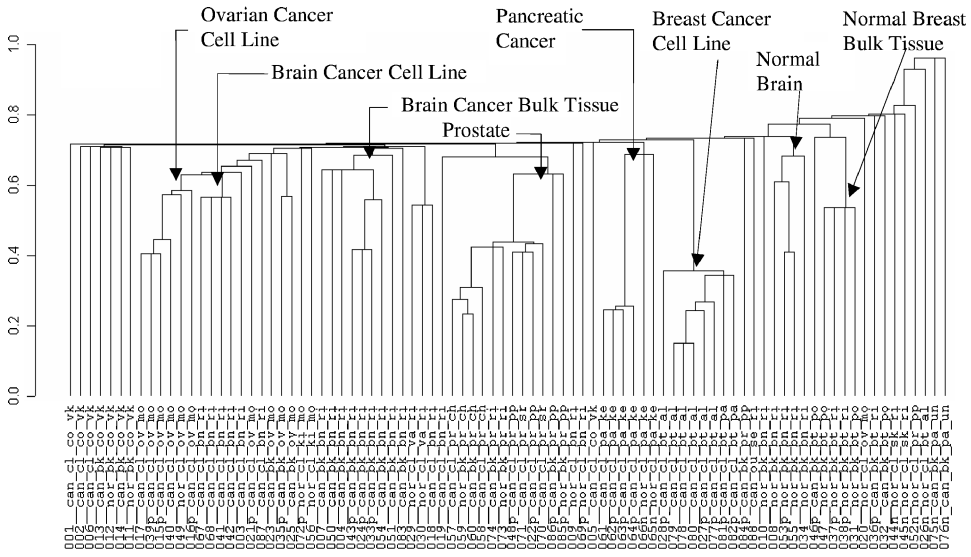


(b) Dedrogram

Fig. 4. Result of OPTICS on the subspace selected by the Wilcoxon test including error removal, normalization, but NO missing tag imputation.



(a) Reachability Plot



(b) Dedrogram

Fig. 5. Result of OPTICS on the subspace selected by the Wilcoxon test including error removal, normalization, and missing tag imputation.

that these two cancers may be related in some way. Furthermore, the ovarian libraries and the brain libraries were produced by two different laboratories. In a later section, we will investigate why these two types of cancers clustered together and which tags they are similar for.

- We also see that we have succeeded to some degree in focusing our data on tags that are related to cancer growth and development. For instance, brain cancer bulk tissue and normal brain bulk tissue have been separated, even though the were produced by the same laboratory (ri). But some further analysis may be needed since the brain cancer bulk tissue and brain cancer cell line clusters are in two different clusters, which means that there are still significant differences between bulk and cell line tissue in the selected subspace.
- We also see a larger breast cancer cell line cluster than in any of our previous results, and a normal brain cluster and a pancreatic cancer cluster have appeared for the first time.

These interesting observations are only possible when all four preprocessing steps—error removal, imputation of missing values, normalization, and subspace selection by the Wilcoxon test—are combined. It is worth emphasizing that the clusters are not only separated according to neoplastic state but also according to tissue type, although we did not utilize any tissue specific information in the subspace selection process—which is important for an unbiased analysis of similarities and dissimilarities of cancers across different tissues.

In this investigation, using the limited resources available on the Web,we did not find any sign of a type of cancer having distinct subtypes. This may, however, be due to the fact that there are not enough libraries of any one type to detect subtypes if they exist. Only when more SAGE libraries for cancerous samples of the different tissue types are available will it be possible to investigate this problem more thoroughly.

The SAGE data clearly contains errors from various sources, each of which can be dealt with individually. The noise created by sequencing error can be repaired by removing ultra-low frequency tags. The problems created by truncated libraries can be reduced through missing value imputation. The discrepancies in the sizes of the various libraries can be dealt with by normalization, and the entire clustering structure can be enhanced and focused by using the Wilcoxon test to select the most relevant tags.

It is also clear at this point that the SAGE method is a valid way to measure gene expression. An interesting point in general is that these libraries were created in different laboratories of various research institutions across North America, and yet the libraries clustered very consistently by tissue type, source and neoplastic state. This is evidence that the SAGE method itself is highly reproducible—an issue that has not been studied very much due to the prohibitive expense of producing duplicated SAGE libraries.

Moreover, it can be seen that SAGE libraries cluster primarily by tissue type. Brain and breast tissues appear to undergo further separation by neoplastic state and tissue source, and ovarian and brain cancer cell lines appear to be more similar to each other than to any other tissue types, while some libraries

do not form into clusters at all. However, it is also clear that more analysis of both the various tissues and the tags will be necessary.

3. FURTHER INVESTIGATIONS

The clustering structure shown in Figure 5 leads us to further questions about the SAGE data. For instance, it would be of interest to explore which tags are significantly different between the various clusters in order to determine what is unique about each cluster. It is also of value to analyze the properties of SAGE libraries made from different tissues, since some tissues formed into more clusters than others, as well as to study those libraries that did not form clusters or are part of a different cluster than expected. Below we describe our findings on two further investigations.

3.1 Identification of Discriminating Genes

Earlier we demonstrated the use of the Wilcoxon test in selecting an appropriate subspace that would accentuate the differences between cancerous and normal tissues. However, since the Wilcoxon test identifies attributes that are significantly different between any two groups, it can also be used to highlight the differences between different clusters.

For example, let us consider the ovarian cancer cell line and brain cancer cell line clusters shown in Figure 5. It would be of interest to determine which tags have different expression levels between the libraries in the brain and ovarian cell line clusters and the normal brain and ovarian libraries (i.e. what transcripts are common to these two cell lines and which discriminate them from their normal counterparts). We expect the tags that passed the Wilcoxon test to represent a subset of genes that allow us to distinguish between normal brain/ovary libraries and the brain/ovarian cancer cell line libraries. We used the UniGene cluster "reliable" mapping provided on the SAGE website, which maps SAGE tags to UniGene Ids that represent nonredundant sets of gene-oriented sequence clusters in GenBank (a molecular sequence database maintained at http://www.ncbi.nlm.nih.gov/). Using this mapping, 165 tags selected by this test were mapped to 169 genes, with 7 tags mapping to 2 genes. Upon examination, some of the genes were excluded from further literature research due to inadequate information. For the remaining genes, we looked up the entries in the OMIM database [Hamosh et al. 2002]. 88 of them did not map to any entries, while 63 of the rest mapped to entries not related to cancer, and 18 mapped to entries related to cancer. Table I shows selected genes that we have associated with a role in cancer through literature research. Columns one, two, and three refer to UniGene cluster identification numbers, their description, and OMIM numbers, respectively. The fourth column summarizes the tag counts for the two types of libraries. For instance, in the second entry, the row <8,0,1,0> indicates that among the 9 normal libraries, 8 do not have the gene expressed, and the remaining one has an expression level between 2 and 5. The PubMed ID is included for a quick reference to the literature used in this study. Furthermore, the last column indicates whether there is general consensus within the literature. All the counts are normalized to the lowest value in

Table I. Information about Selected Genes

UniGene Cluster	UniGene Cluster Description	OMIM Number	Expression					PubMed ID	Lit.
				0	**1**	**≥2**	**≥5**		
Hs.179565	MCM3 Minichromosome Maintenance Deficient 3	602693	**N**	4	2	3	0	11801723 10653597	√
			C	1	1	5	7		
				0	**1**	**≥2**	**≥5**		
Hs.334562	CDC2 Cell Division Cycle 2	116940	**N**	8	0	1	0	11836499	√
			C	3	1	9	1		
				0	**1**	**≥2**	**≥5**		
Hs.241257	LTBP1 Latent Transform. Growth Factor Beta Binding Protein 1	150390	**N**	6	2	1	0	11376559	√
			C	2	0	7	5		
				0	**1**	**≥2**	**≥5**		
Hs.239	FOXM1 Foxhead Box M1	602341	**N**	8	0	1	0	11682060	√
			C	1	2	5	6		
				0	**1**	**≥2**	**≥5**		
Hs.45231	LDOC1 Leucine Zipper, down regulated by cancer 1 (AAGGTGGCAT)	n/a	**N**	5	3	0	1	10403563	×
			C	3	1	9	1		

the table. Due to space limitations, we include only five genes in the following discussion.

The first entry is one of the discriminating genes identified—the *minichromosome maintenance deficient 3* (MCM3) gene. The MCM3 protein is involved in eukaryotic genome replication and participates in the regulation of DNA duplication. This has been used as a cell growth and differentiation marker for cancer prognosis. Although MCM3 is also expressed in nonproliferating cells as described in the reference paper, only five of nine normal libraries in our data expressed MCM3, whereas thirteen of fourteen cancer cell line libraries expressed MCM3 with a high tag count.

The second entry is CDC2. Together with BRCA1, CDC2 is involved in modulating the cell cycle arrest process, specifically at the G2/M checkpoint [Yarden et al. 2002]. In the OMIM entry ID116940, the CDC2 protein is proposed to be involved in the resistance of drug-induced cell death in breast cancer due to its role in the regulation of the cell division cycle. Of the fourteen brain/ovarian cancer cell line libraries, eleven of them expressed CDC2, ten more than the normal libraries.

For the third entry, in a study of ovarian cancer using the differential display method, Higashi et al. [2001] identified the LTBP1 transcript as one of the most highly differentially expressed transcripts in both ovarian cancer tissue and cell line material. In other malignant tissues LTBP genes have been associated with reduced expression [Oklu and Hesketh 2000] and elevated levels in their surrounding extracellular matrix and stromal tissues. This helps confirm our analysis that LTBP1 is a discriminating gene in resolving between ovarian cancer, normal tissues and other malignancies as well as indicating that the LTBP1 is also upregulated in brain cancers.

FOXM1 (previously known as Trident) is a transcription factor, expressed in many cell types undergoing proliferation [Leung et al. [2001]. Its proposed

function is within the process of cell cycle regulation through the control of expression of Cyclin B1. Thirteen cancer cell line libraries (8/8 brain & 5/6 ovary) showed elevated levels of FOXM1 transcripts compared to only one of the normal libraries. Furthermore, three libraries that expressed FOXM1 were also found to express Cyclin B1 but not Cyclin D1, as documented by Leung et al. [2001].

All of these are examples showing that our methodology seems to produce relevant results. In fact, there is little in the literature that directly links the genes to brain and ovarian cancer. Our approach serves to establish previously undetermined linkages and commonalities between brain and ovarian cancer.

We note that not every gene we found agrees with the previously established literature. For example LDOC1 (leucine zipper down regulated in cancer) was identified as being ubiquitously expressed in normal tissues and down-regulated in cancer cell lines by Nagasaki et al. [1999]. Our examination points to a different conclusion. In fact, analysis aside, LDOC1 is not ubiquitously expressed in our normal brain/ovary SAGE libraries as claimed in the paper. Instead of being down regulated in various tissues, our observed tag counts were significantly higher in brain/ovary cancer cell line libraries when compared to their normal counterparts. However, one aspect of the SAGE technology is that different transcripts may produce similar tags. Therefore it is possible that this particular tag also derives from a currently undiscovered transcript, although the relatively high expression cell line material suggests that it would be relatively well represented in existing resources of ESTs (Expressed Sequence Tags) and cDNAs (complementary copy of an mRNA).

3.2 Gene Expression Classification

The analysis just presented provides a method for identifying genes that discriminate one group of libraries from another. The SAGE libraries also present an opportunity for us to consider a related question: whether different cancerous tissues have different identities at the gene expression level. To evaluate this question, we adopt a classification approach. For each case, we randomly selected 50 to 90 percent of normal and cancerous libraries of a specific tissue (e.g. brain) and labelled them as the training libraries. Then we randomly picked a testing library from the remaining ones. The goal was to try to predict whether the testing library is cancerous or normal. Due to the uneven number of libraries in each state (i.e. normal/cancerous cell line/bulk) in the four tissues, it was fairer just to classify the libraries according to their neoplastic state and disregard whether they are bulk tissue or cell line. The dissimilarities between the testing library and the training libraries were measured using correlation coefficient, and the closest training library was identified. The testing library was then predicted to be cancerous or normal, depending on whether the closest training library was cancerous or normal. We repeated this procedure a hundred times for different randomizations, and recorded the percentage of times the correct prediction was made. The results are shown in Figure 6.

Both brain and breast cancer libraries have high prediction accuracy, suggesting that these two cancer types have strong identities. Furthermore, the

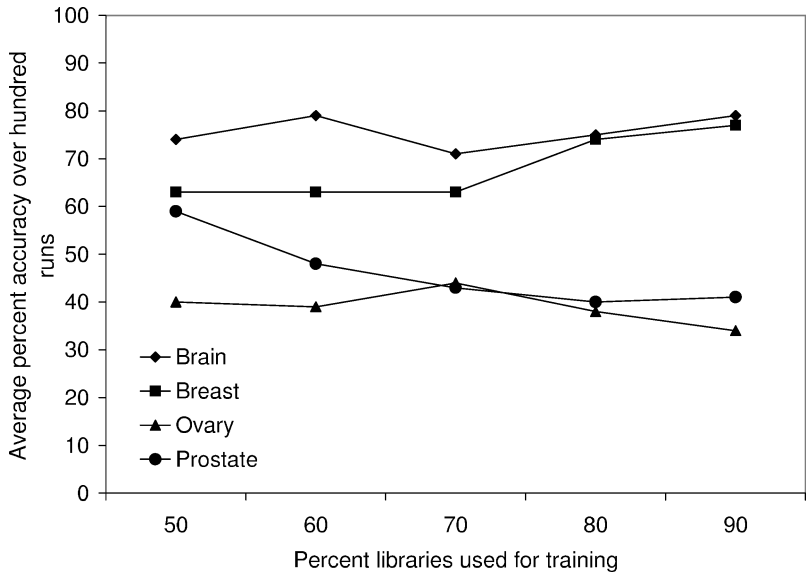


Fig. 6. Nearest Neighbour Classification with Random Sampling using correlation coefficient. The x-axis of the graph shows the percentage of libraries used as training libraries. The y-axis shows the average percent accuracy over a hundred runs.

more libraries used for training, the higher the accuracy. In contrast, both prostate and ovarian cancer libraries have low accuracy. Specifically, the number of normal and cancerous prostate libraries are 4 and 10 respectively. In this case, always predicting cancer would already give an accuracy of 71% (10 out of 14). The curve in Figure 6 for prostate indicates that the training libraries serve little purpose. This suggests that normal prostate and cancerous prostate libraries are very similar to each other, and there seems to be a lack of a definitive signature of prostate cancer. As a consequence, the increased number of training libraries did not improve the performance, but indeed confused matters further. This may explain the inverse relationship between the percentage of training libraries used and the accuracy of the prediction. A similar comment applies to ovarian tissue.

3.3 Outliers

In general, the clustering structure shown in Figure 5 is quite strong, and there are very few instances of mixed clusters or outliers. The majority of outliers are libraries of tissue types of which there are only 2 instances, such as kidney and skin tissue libraries. It is interesting to note that there are 7 colon tissue libraries which were all closer to each other than to any other tissue type but did not form a pronounced cluster. This may be because the colon tissue libraries were not all from the same tissue source, or perhaps because colon tissues vary more in gene expression from person to person than other tissues.

It is also interesting to note that the 6 libraries that were the furthest from all other libraries in the data set were all truncated, which meant that

there were no complete libraries available to perform missing tag imputation. This is additional evidence that the amount of data in a library affects how it clusters.

The only true cases of libraries of different tissues forming a single cluster are the ovarian and brain cancer cell line cluster and the vascular and brain cluster. In the former case, the two tissues form sub-clusters of a larger combined cluster as previously discussed. In the latter case, the two vascular tissue libraries form a small cluster with a brain cancer cell line library. However, with only three libraries in total, the significance of this cluster is not clear. Because the brain cancer library is not close to the other brain cancer libraries, it is possible that this library represents a subtype of brain cancer that is similar to vascular tissue rather than to ovarian cancer tissue.

4. CONCLUSION

In this article we proposed a method for clustering SAGE data to detect similarities and dissimilarities between different types of cancer at the sub-cellular level. We introduced four preprocessing steps to reduce errors, restore missing data, normalize the data, and select an appropriate subspace based on the Wilcoxon test. After all of these steps were performed, the clustering algorithm OPTICS produced a promising hierarchical clustering structure in which the SAGE libraries are grouped according to tissue type and neoplastic state, showing a possible relationship between brain and ovarian cancer.

We have shown that the SAGE data appears to be highly reproducible, since the data was produced in various different laboratories but still displays a consistent clustering structure, when appropriate preprocessing steps were applied. We have also shown that the Wilcoxon test can be used to identify discriminating tags. Furthermore, we have shown that different cancerous tissues have different degrees of identity at the gene expression level. This may suggest that the strong-identity ones, like brain and breast, may be good candidates for applying the kind of discriminating gene identification described here rather than the weak-identity ones.

The dissimilarities between the libraries in the subgroups of several tissue types were studied. Brain tissues were found to be completely separated by subgroup—every brain tissue library of every subgroup was found to be closest to other brain libraries of the same subgroup. One breast cancer bulk tissue library was found to be closer to a normal breast bulk tissue library than to other breast cancer tissue libraries, perhaps because of contamination by surrounding cells in the sample. Prostate tissue library subgroups were all close to each other, suggesting that gene expression in cancerous prostate tissues is not very different from that in normal tissues. Normal ovarian cell line tissues were found to be close to ovarian cancer cell line tissues, which may mean that normal cell lines may be used as a model for cancer in ovarian tissue.

This is, however, preliminary work in several respects. Further research is necessary to investigate the significance of our findings, and more data is required before subtypes of cancers can be discovered. We based our analysis on 88 SAGE libraries, but more libraries are added over time. However, SAGE data

is expensive to produce and it may take some time before a large enough data set is available. In addition, more aggressive methods for subspace selection could be explored to reveal more subtle similarities between different cancers and dissimilarities within a specific cancer. Nevertheless, producing more SAGE data is important because, as our experiments indicate, it contains valuable information that is not available elsewhere.

The main purpose of this analysis was to identify cancer types and subtypes, which is why we focused on clustering only the SAGE libraries. In the future it may also be interesting to cluster the SAGE tags as well. Understanding the clustering structure of the tags will not be trivial, however, because of the large number of tags and the fact that the majority of tags have not yet been mapped to a gene.

ACKNOWLEDGMENTS

We would like to thank the anonymous TOIS reviewers and the editor whose constructive comments contributed greatly to clarity and precision of this article.

REFERENCES

ALIZADEH, A. A., EISEN, M. B., DAVIS, R. E., MA, C., LOSSOS, I. S., ROSENWALD, A., BOLDRICK, J. C., SABET, H., TRAN, T., YU, X., POWELL, J. I., YANG, L., MARTI, G. E., MOORE, T., HUDSON, J., LU, L., LEWIS, D. B., TIBSHIRANI, R., SHERLOCK, G., CHAN, W. C., GREINER, T. C., WELSENBURGER, D. D., ARMITAGE, J. O., WARNKE, R., LEVY, R., WILSON, W., GREVER, M. R., BYRD, J. C., BOTSTEIN, D., BROWN, P. O., AND STAUDT, L. M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 3 (Feb.), 503–511.

ALON, U., BARKAI, N., NOTTERMAN, D. A., GISH, K., YBARRA, S., MACK, D., AND LEVINE, A. J. 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.

ANKERST, M., BREUNIG, M., KRIEGEL, H.-P., AND SANDER, J. 1999. OPTICS: Ordering Points to identify the clustering structure. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Philadelphia, PA, June 1999, ACM Press, New York, NY, 49–60.

BEN-DOR, A., SHAMIR, R., AND YAHKINI, Z. 1999. Clustering gene expression patterns. *J. Comput. Biol. 6*, 281–297.

BEN-DOR, A., BRUHN, L., FRIEDMAN, N., NACHMAN, I., SCHUMMER, M., AND YAKHINI, Z. 2000. Tissue classification with gene expression profiles. *J. Comput. Biol. 7*, 559–584.

BOON, K., OSÓRIO, E. C., GREENHUT, S. F., SCHAEFER, C. F., SHOEMAKER, J., POLYAK, K., MORIN, P. J., BUETOW, K. H., STRAUSBERG, R. L., DE SOUZA, S. J., AND RIGGINS, G. J. 2002. An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. USA 99*, 11287–11292 .

BUCKHAULTS, P., ZHANG, Z., CHEN, Y. C., WANG, T. L., ST CROIX, B., SAHA, S., BARDELLI, A., MORIN, P. J., POLYAK, K., HRUBAN, R. H., VELCULESCU, V. E., AND SHIH, IEM. 2003. Identifying tumor origin using a gene expression-based classification map. *Cancer Res. 15*, 63, 14, 4144–4149.

EDGAR, R., DOMRACHEV, M., AND LASH, A. E. 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res. 30*, 207–210.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O., AND BOTSTEIN, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA 95*, 25, 14863–14868.

GRAY, J. W. AND COLLINS, C. 2000. Genome changes and gene expression in human solid tumors. *Carcinogenesis 21*, 443–52.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D., AND LANDER, E. S. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

HAMOSH, A., SCOTT, A. F., AMBERGER, J., BOCCHINI, C., VALLE, D., AND MCKUSICK, V. A. 2002. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res. 30*, 52–55.

HAN, J. AND KAMBER, M. 2000. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, San Francisco, CA.

HASHIMOTO, S., NAGAI, S., SESE, J., SUZUKI, T., OBATA, A., SATO, T., TOYODA, N., DONG, H. Y., KURACHI, M., NAGAHATA, T., SHIZUNO, K., MORISHITA, S., AND MATSUSHIMA, K. 2003. Gene expression profile in human leukocytes. *Blood 101*, 9, 3509–3513.

HIGASHI, T., SASAGAWA, T., INOUE, M., OKA, R., SHUANGYING, L., AND SAIJOH, K. 2001. Overexpression of latent transforming growth factor-beta 1 (TGF-beta 1) binding protein (LTBP-1) in association with TGF-beta 1 in ovarian carcinoma. *Jpn. J. Cancer Res. 92*, 2, 506–515.

LAL, A., LASH, A. E., ALTSCHUL, S. F., VELCULESCU, V., ZHANG, L., MCLENDON, R. E., MARRA, M. A., PRANGE, C., MORIN, P. J., POLYAK, K., PAPADOPOULOS, N., VOGELSTEIN, B., KINZLER, K. W., STRAUSBERG, R. L., AND RIGGINS, G. J. 1999. A public database for gene expression in human cancers. *Cancer Res. 59*, 5403–5407.

LASH, A. E., TOLSTOSHEV, C. M., WAGNER, L., SCHULER, G. D., STRAUSBERG, R. L., RIGGINS, G. J., AND ALTSCHUL, S. F. 2000. SAGEmap: A public gene expression resource. *Genome Res, 10*, 7, 1051–1060.

LEUNG, T. W., LIN, S. S., TSANG, A. C., TONG, C. S., CHING, J. C., LEUNG, W. Y., GIMLICH, R., WONG, G. G., AND YAO, K. M. 2001. Over-expression of FoxM1 stimulates cyclin B1 expression. *FEBS Lett. 507*, 59–66.

NACHT, M., DRACHEVA, T., GAO, Y., FUJII, T., CHEN, Y., PLAYER, A., AKMAEV, V., COOK, B., DUFAULT, M., ZHANG, M., ZHANG, W., GUO, M., CURRAN, J., HAN, S., SIDRANSKY, D., BUETOW, K., MADDEN, S. L., AND JEN, J. 2001. Molecular characteristics of non-small cell lung cancer. *Proc. Natl. Acad. Sci. USA. 98*, 26, 15203–15208.

NCBI (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION) SAGE: Measuring Gene Expression, http://www.ncbi.nlm.nih.gov/SAGE.

NAGASAKI, K., MANABE, T., HANZAWA, H., MAASS, N., TSUKADA, T., AND YAMAGUCHI, K. 1999. Identification of a novel gene, LDOC1, down-regulated in cancer cell lines. *Cancer Lett. 140*, 227–234.

NG, R. T. AND HAN, J. 1994. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, September 1994, Morgan Kaufmann Publishers, San Francisco, CA, 144–155.

NG, R. T., SANDER, J., AND SLEUMER, M. 2001. Hierarchical cluster analysis of SAGE data for cancer profiling. *Workshop on Data Mining in Bioinformatics*. In *Conjunction with 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, August 2001.

OKLU, R. AND HESKETH, R. 2000. The latent transforming growth factor beta binding protein (LTBP) family. *Biochem J. 352*, Pt 3, 601–610.

PEROU, C. M., JEFFREY, S. S., VAN DE RIJN, M., REES, C. A., EISEN, M. B., ROSS, D. T., PERGAMENSCHIKOV, A., WILLIAMS, C. F., ZHU, S. X., LEE, J. C. F., LASHKARI, D., SHALON, D., BROWN, P. O., AND BOTSTEIN, D. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Natl. Acad. Sci USA 96*, 9212–9217.

PORTER, D. A., KROP, I. E., NASSER, S., SGROI, D., KAELIN, C. M., MARKS, J. R., RIGGINS, G., AND POLYAK, K. 2001. A SAGE (serial analysis of gene expression) view of breast tumor progression. *Cancer Res. 61*, 15, 5697–702.

SANDER, J., QIN, X., LU, Z., NIU, N., AND KOVARSKY, A. 2003. Automatic extraction of clusters from hierarchical clustering representations. In *Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Seoul, Korea, April/May 2003. Lecture Notes in Artificial Intelligence 2637, Springer, Berlin, Germany, 75–87.

STOLLBERG, J., URSCHITZ, J., URBAN, Z., AND BOYD, C. D. 2000. A Quantitative Evaluation of SAGE. *Genome Res. 10*, 1241–1248.

STRAUSBERG, R. L, BUETOW, K. H., EMMERT-BUCK, M. R., AND KLAUSNER, R. D. 2000. The cancer genome anatomy project: Building an annotated index. *Trends Genet. 16*, 3, 103–106.

TANNER, M. M., GRENMAN, S., KOUL, A., JOHANNSSON, O., MELTZER, P., PEJOVIC, T., BORG, Å., AND ISOLA, J. J. 2000. Frequent Amplification of Chromosomal Regoin 20q12-q13 in Ovarian Cancer. *Clin. Cancer Res. 6*, 1833–1839.

TAVAZOIE, S, HUGHES, J. D., CAMPBELL, M. J., CHO, R. J., AND CHURCH, G. M. 1999. Systematic determination of genetic network architecture. *Nature Genetics 22*, 281–285,

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P. HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D., AND ALTMAN, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics 17*, 6, 520–525.

VAN RUISSEN, F., JANSEN, B. J., DE JONGH, G. J., VAN VLIJMEN-WILLEMS, I. M., AND SCHALKWIJK, J. 2002. Differential gene expression in premalignant human epidermis revealed by cluster analysis of serial analysis of gene expression (SAGE) libraries. *FASEB J. 16*, 2, 246–248.

VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B., AND KINZLER, K. W. 1995. Serial analysis of gene expression. *Science 270*, 484–487.

WILCOXON, F. 1945. Individual Comparisons by Ranking Methods. *Biometrics 1*, 80–83.

YARDEN, R. I., PARDO-REOYO, S., SGAGIAS, M., COWAN, K. H., AND BRODY, L. C. 2002. BRCA1 regulates the G2/M checkpoint by activating Chk1 kinase upon DNA damage. *Nature Genetics 30*, 285–289.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E., AND RUZZO, W. L. 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics 17*, 977–987.

ZHANG, L., ZHOU, W., VELCULESCU, V. E., KERN, S. E., HRUBAN, R. H., HAMILTON, S. R., VOGELSTEIN, B., AND KINZLER, K. W. 1997. Gene expression profiles in normal and cancer cells. *Science 276*, 1268–1272.

Received October 2003; revised June 2004, August 2004; accepted August 2004