Cancer Genomics and Molecular Pattern Recognition

Pablo Tamayo [*] & Sridhar Ramaswamy [*] [†]

[*] Cancer Genomics Group, Whitehead Institute / Massachusetts Institute of Technology
Center for Genome Research, Cambridge, MA 02139; [†] Department of Adult Oncology,
Dana-Farber Cancer Institute / Harvard Medical School, Boston, MA 02115

June 23, 2002

Contents:

Cancer Genomics

Cancer is a genetic malady, mostly resulting from acquired mutations and epigenetic changes that influence gene expression. Accordingly, a major focus in cancer research is identifying genetic markers that can be used for precise diagnosis or therapy. Over the last half-century, investigators have used reductionism to discover such markers through the study of simple genetic changes like balanced chromosomal translocations. For example, fundamental insights into the nature of the *bcr-abl* gene translocation product resulted in the precise molecular classification of chronic myelogenous leukemia and recently led to the development of the molecularly targeted tyrosine kinase inhibitor STI571 (Gleevec; Novartis, East Hanover, NJ) for the treatment of this disease. Ninety percent of human cancers, however, are epithelial in origin and display marked aneuploidy, multiple gene amplifications and deletions, and genetic instability, making resulting downstream effects difficult to study with traditional methods. Because this complexity probably explains the clinical diversity of histologically similar tumors, a comprehensive understanding of the genetic alterations present in all tumors is required.

The initial sequencing of the human genome, coupled with technologic advances, now make it possible to embrace the genetic complexity of common human cancers in a global fashion. Tools are currently available, or are being developed, for the identification of all changes that take place in cancer at the DNA, RNA, and protein levels. In particular, the use of DNA microarrays for the comprehensive analysis of RNA expression (expression profiling) in human tumor samples holds much promise (see review articles in the Chipping Forecast 1999).

A major challenge with this approach, however, remains the interpretation of complex and biologically “noisy” data in a way that yields new knowledge. We have therefore focused on developing first-generation approaches to gene expression data analysis that are suitable for this purpose. Without such analytic tools, DNA microarray data are useless. This chapter is meant to serve as an introduction to fundamental concepts and techniques that have been developed in gene expression data mining over the last three years. It is not meant to be a comprehensive review of this rapidly expanding field, nor is it a step-by-step set of recipes. Most of the examples described come from our experience in cancer gene expression data analysis at the Whitehead / MIT Center for Genome Research over the last five years, but references to other works are also given when relevant to the discussion.

Basic Data Analysis

Tumors are heterogeneous mixtures of different cell types, including malignant cells with varying degrees of differentiation, stromal elements, blood vessels, and inflammatory cells. Two tumors with similar clinical stages can vary markedly in grade and in relative proportions of different elements (e.g., prostatic adenocarcinoma). Tumors of different grades might potentially differ in gene expression, and different markers can be expressed either by malignant cells or by other cellular elements. Because this heterogeneity can complicate the interpretation of gene expression studies, sample selection is an important issue that must be kept in mind when analyzing tumor gene expression data.

Multiple sources of variation that must be understood in evaluating any microarray experiment include the following: (1) varying cellular composition among tumors, (2) genetic heterogeneity within tumors due to selection and genomic instability, (3) differences in sample preparation, (4) nonspecific cross-hybridization of probes, and (5) differences between individual microarrays. In general, biologic variation is the major source of variation in gene expression experiments. Increasing the sample number can help in understanding the range of biologic variation in an experiment. Variation due to technical factors can be addressed by replicating sample preparation or array hybridization. Although most high throughput expression profiling centers have informal criteria for what constitutes bad data, however, there are no generally accepted guidelines. For approaches to microarray experimental design and the analysis of variation see Cheng and Wong 2001a, Tseng et al 2001, Hunter et al 2001, Kerr A. and G. Churchill 2001a,b).

Basic data analysis consists of preparing datasets for higher-level analysis such as clustering or class prediction. This “pre-processing” of raw data can have profound effects on subsequent analysis and has to be done by considering the idiosyncrasies of the original gene expression technology platform (i.e. “chip type”). For example, cDNA microarrays generate gene ratio data between fluorescence intensities of experimental and control samples on a gene-by-gene basis. In contrast, oligonucleotide microarrays such as the Affymetrix GeneChip platform generate absolute expression values from a single sample. Each microarray platform generally has software packages that provide one “file” per sample containing one gene per “row.” These sample files are usually combined into multi-sample files for further analysis. Our discussion of data analysis starts at this point.

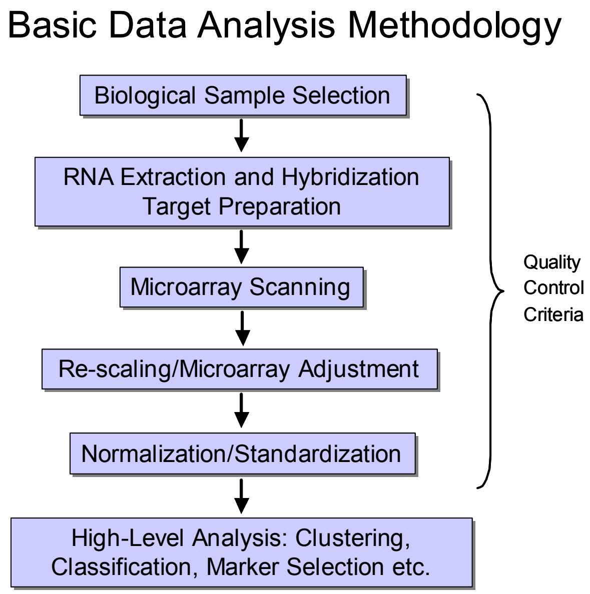

Basic Data Analysis Methodology
Biological Sample Selection
RNA Extraction and Hybridization Target Preparation
Microarray Scanning
Re-scaling/Microarray Adjustment
Normalization/Standardization
High-Level Analysis: Clustering, Classification, Marker Selection etc.
Quality Control Criteria

Figure 1. Methodology for basic data analysis.

Raw Data Quality Control

The quality of each microarray profile is generally assessed using measurements of overall microarray fluorescence intensity (e.g. mean, variance), the distribution of feature or spot intensities, and the proportion of total genes receiving significant signal. Any microarray that fails these quality control measures is generally excluded from downstream analysis. Replicate experiments for each sample can be used to focus on those gene measurements with the highest reproducibility (Lee et al 2000, Kerr et al 2001). With technologic improvements, however, raw data quality is presently quite good in experienced hands. Therefore, we currently emphasize the analysis of larger numbers of samples rather than studying fewer samples and more replicates.

Scaling

Raw gene expression data from multiple samples ("chips") is generally scaled to compensate for global differences in chip intensities and microarray to microarray variation. This can be done using simple multiplicative factors to match overall mean intensities among microarrays. Other more sophisticated methods use model-based approaches to compensate for probe-specific biases (Cheng and Wong 2001b).

Thresholding, Filtering, & Normalization

In some cases it may be desirable to threshold and ceiling the data, since very low and very high microarray fluorescence readings are less reliable and reproducible. As many clustering and classification algorithms work better with smaller number of genes, or are especially sensitive to noisy profiles, genes that show low or flat expression across multiple samples are usually filtered out of datasets. One of the simplest ways to do this is by using a variation filter which tests for a minimum fold-change (max / min) and absolute variation (max - min) among samples and excludes genes not passing the corresponding thresholds. The precise parameters of variation filters are problem-, dataset- and platform-dependent and different thresholds and stringencies in the variation filter may be used depending on the particular analysis. After filtering, and before higher-level data analysis, one may also consider normalizing each gene to a mean of 0 and variance of 1 across all samples. This strategy can be useful if one is interested in emphasizing relative rather than absolute differences in gene intensity.

Higher-Level Data Analysis: Unsupervised & Supervised Learning

To date, the higher-level computational analysis of gene expression data has centered on two approaches (Golub et al 1999). Unsupervised learning, or clustering, involves the aggregation of a diverse collection of data into clusters based on different features in a data set. For example, one could divide a group of people into clusters based on any combination of eye color, waist size, or height. Similarly, one can gather data about the various expressed genes in a collection of tumor samples and then cluster the samples as best as possible into groups based on the similarity of their aggregate expression profiles. Alternatively, one could cluster genes across all samples, to identify genes that share similar patterns of expression in varying biologic contexts. Such approaches have the advantage of being unbiased and allow for the identification of structure in a complex data set without making any a priori assumptions. However, because many different relationships are possible in a complex data set, the predominant structure uncovered by clustering may not necessarily reflect clinical or biologic distinctions of interest.

In contrast, supervised learning incorporates the knowledge of class label information to make distinctions of interest. A training data set is used to select those features that best make a distinction. These features are then applied to an independent test data set to validate the ability of selected features to make that distinction. For example, one could select a subset of expressed genes that are best able to distinguish between two cancer types and build a computational model that uses these selected genes to sort an independent, unlabelled collection of those tumor types into the two groups of interest. However, supervised learning is dependent on accurate sample labels, which can be an issue given the limitations of histopathologic cancer diagnosis. Sometimes, results from unsupervised and supervised learning on a single data set can overlap, but this does not have to be the case.

An important issue with either analytic approach is that of statistical significance of observed correlations. A typical microarray experiment yields expression data for thousands of genes from a relatively small number of samples, and gene-class correlations, therefore, can be revealed by chance alone. This issue can be addressed by collecting more samples for each class studied, but this is often difficult with clinical cancer samples. Another approach is to perform exploratory data analysis on an initial data set and apply findings to an independent test set. Findings confirmed in this fashion are less likely a result of chance. Permutation testing, which involves randomly permuting class labels and determining gene-class correlations, has also been used to determine statistical significance (Golub et al 1999). Observed gene-class correlations that are stronger than those seen in permuted data are considered statistically significant.

Unsupervised Learning: Clustering

In unsupervised learning techniques, the structure in a data set is elucidated without using any a priori assumptions or knowledge as part of exploratory data analysis. The promise of these methods lies in their ability to provide a molecular grouping or taxonomy of samples or genes. One of the easiest ways to analyze data in this context is by using a clustering algorithm (Hartigan 1975, Gordon 1981, Duda et al 2000). Objects of interest, usually genes or samples, are classified into groups according the how "close" they are to each other. This is accomplished by using a "distance," correlation, or "similarity" function in the clustering algorithm. For example, one can cluster a set of biological samples by their Euclidean distances by considering all gene expression values in a dataset:

$$\text{Distance(sample x, sample y)} = \sqrt{}((E^{x}_{\text{gene 1}} - E^{y}_{\text{gene 1}})^2 + (E^{x}_{\text{gene 2}} - E^{y}_{\text{gene 2}})^2 + \ldots)$$

Here $E^{x}_{\text{gene 1}}$ is the expression value of gene 1 in the array corresponding to sample x. A clustering algorithm uses these distances to group samples or genes, and it returns an organization scheme to classify them (e.g. a set of clusters or a tree).

Unsupervised learning approaches such as clustering can be very useful when the underlying structure of the data is unknown; however, they have the disadvantage, if unguided, of sometimes producing results that may or may not be relevant to distinctions in the data that are biologically relevant. Clustering often rediscovers already known subclasses or differences if these distinctions are predominant (e.g. estrogen receptor positive versus negative breast cancers). However, this approach can also discover unanticipated relationships, and clustering methods have been used with relative success in a number of cancer classification problems. In practice, it is often challenging to interpret clusters that result from unsupervised learning in cancer datasets. A general methodology for clustering is shown in figure 2.

Some of the first work using this approach in analyzing gene expression involved time series data. Genes were grouped, or clustered, according to their behavior over time, first by eye (Cho et al. 1998) and then by an automated hierarchical technique (Eisen et al. 1998). Hierarchical clustering is an unsupervised learning method useful for dividing data into natural groups by organizing the data into a hierarchical tree structure ("dendogram") based upon the degree of similarity between either samples or genes (Eisen et al 1998). The lengths of branches in a dendogram reflect degree of relatedness. By examining dendogram branches, previously unanticipated relationships between samples and genes can be discovered in a gene expression dataset. Tamayo et al (1999) introduced the use of self-organizing maps (SOMs) for unsupervised learning in the HL-60 model of leukemia differentiation, and found that resulting gene clusters corresponded to pathways involved in the "differentiation" treatment of acute promyelocytic leukemia (APL). The Self Organizing Map (SOM) is a clustering algorithm where a grid of 2D nodes (clusters) is iteratively adjusted to reflect the global structure in the expression dataset (Tamayo et al 1999). With the SOM, the geometry of the grid is randomly chosen (e.g., a

3 x 2 grid) and mapped to the k-dimensional gene expression space. The mapping is then iteratively adjusted to reflect the natural structure of the data. Resulting clusters are organized in a 2D grid where similar clusters lie near to each other and provide an automatic "executive" summary of the dataset.

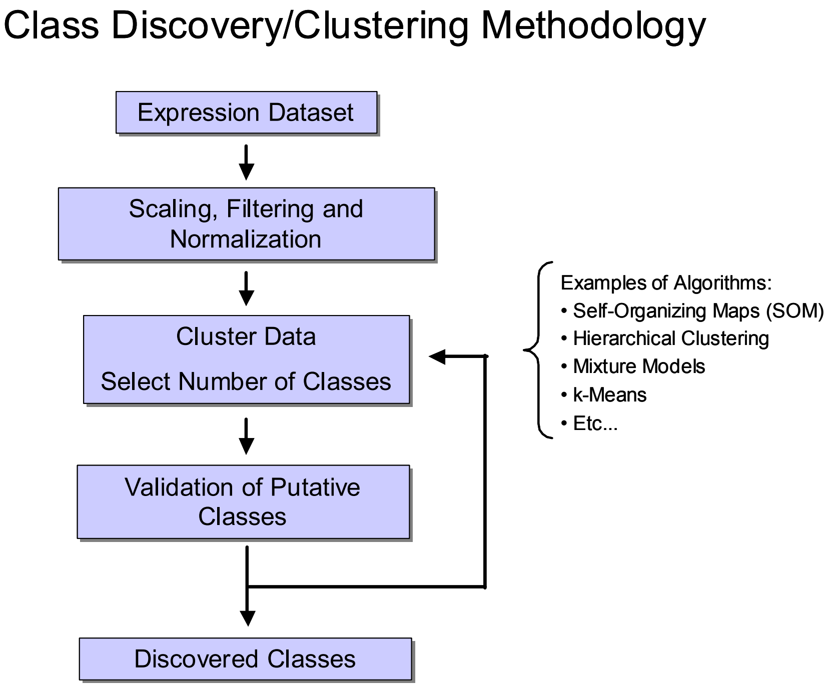

Class Discovery/Clustering Methodology
Expression Dataset
Scaling, Filtering and Normalization
Cluster Data
Select Number of Classes
Validation of Putative Classes
Discovered Classes
Examples of Algorithms:
• Self-Organizing Maps (SOM)
• Hierarchical Clustering
• Mixture Models
• k-Means
• Etc...

Figure 2. General methodology for clustering gene expression data.

Golub et al. used a 2-cluster SOM to automatically cluster an initial set of 38 leukemia samples into two classes based on the expression pattern of 6817 genes (figure 3). They then compared these SOM clusters to the known lymphoblastic vs myeloid leukemia (AML / ALL) distinction. As demonstrated, the two SOM clusters closely paralleled this morphological distinction with the first cluster containing mostly ALLs (24 out of 25 samples) and the second containing mostly AMLs (10 out of 13 samples). Thus, the clustering algorithm was effective but not perfect at separating samples into biologically meaningful groups.

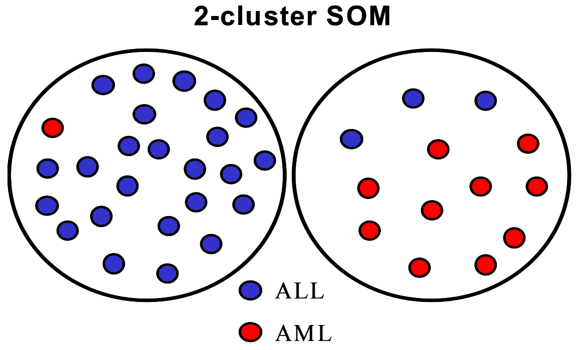


Figure 3. Clustering of Leukemia samples into two groups using a 2X1 SOM.

Golub et al also searched for further sub-classifications of the leukemia samples by constructing a 4-class (2x2) SOM (figure 4). The clustering algorithm was successful at separating the samples into more refined groups reflecting another important biological distinction: different ALL cell lineages (B- and T-Cell).

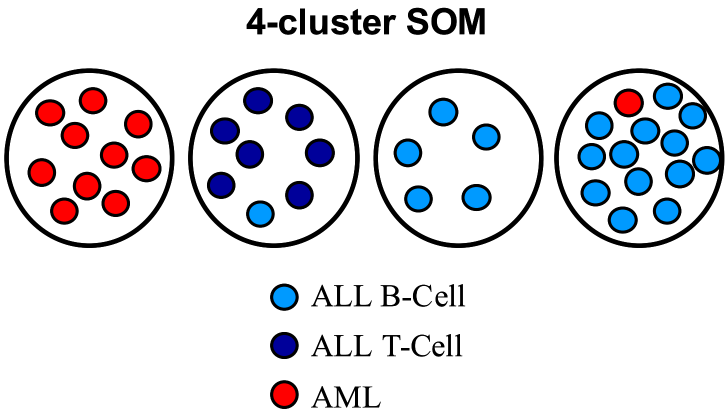


Figure 4. Clustering of Leukemia samples into 4 groups using a 4x1 SOM.

Hierarchical clustering (Eisen et al 1998) was also applied to the same dataset (figure 5). Again, this clustering approach revealed three major leukemia sub-groups, suggesting that robust gene expression differences between different tumor sub-types can be discovered using unsupervised learning.

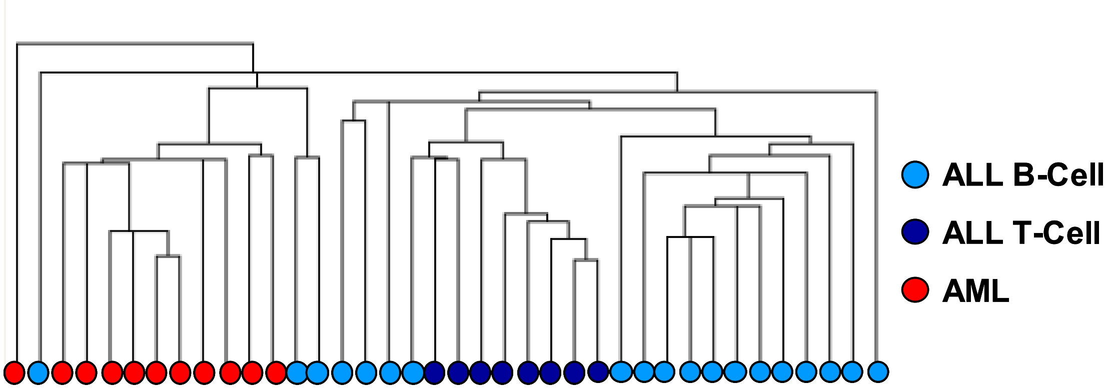


Figure 5. Hierarchical clustering of Leukemia Samples based on the expression of the 330 most varying genes.

Similar studies have recently been described for the sub-classification of various tumor types including breast cancer (Perou et al. 1999; Perou et al. 2000), lung cancer (Bhattacharjee et al 2001) and melanoma (Bittner et al. 2000).

Clustering has yielded results that are interpretable in the context of *a priori* knowledge (i.e. known leukemia sub-classes). However, in the absence of such knowledge the biological interpretation of clustering results remains a challenge. Often clustering results are not in themselves the desired results but the starting point for further interpretation or experimentation.

An area of active research, moreover, involves the statistical interpretation of clustering results. Often asked questions include what constitutes a cluster and what is the statistical significance of a given clustering result? There are presently no good general answers for these important questions, although some groups have proposed the implementation of formal measures of clustering significance such as the gap statistic (Tibshirani 2001).

Supervised Learning: Prediction

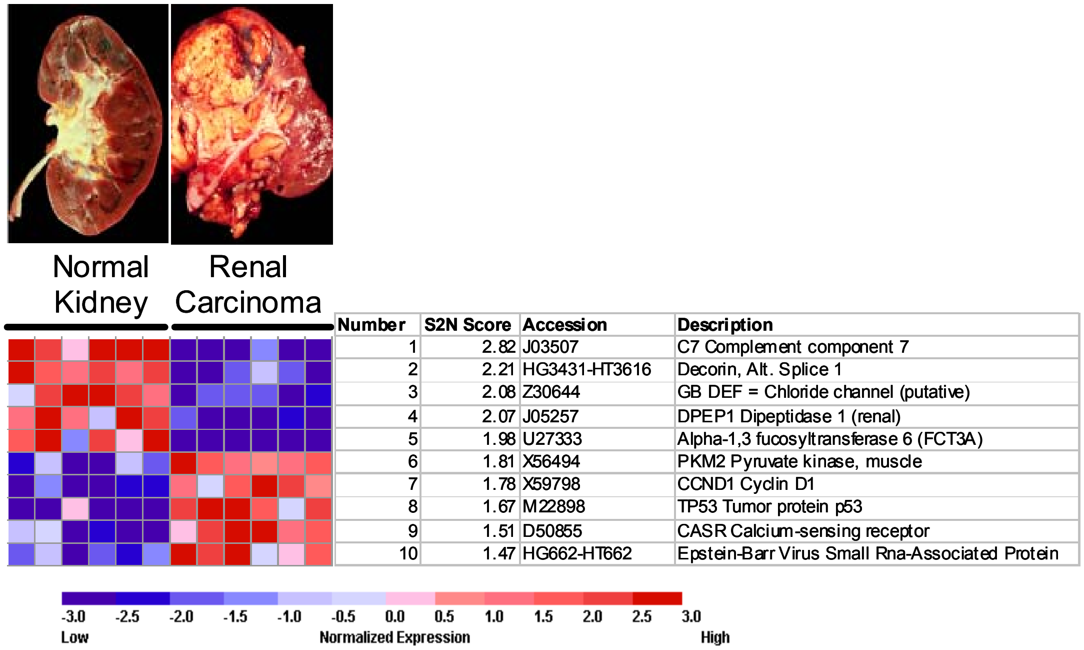
Supervised learning or class prediction methods represents another important paradigm in molecular classification and pattern recognition. The simplest analysis involves selecting the features (genes) most correlated with a phenotypic distinction of interest. These features or "marker genes" are biologically interesting in themselves but they can also be used as the input of a classification algorithm that uses existing "labeled" samples to build a model to predict the labels for future samples. For example, marker genes in a cancer dataset can be fed into a computational classifier to distinguish cancer types on the basis of site / cell of origin or clinical outcome. This powerful approach, supervised machine learning or class prediction (Duda et al 2000, Fukunaga 1990), involves data collection, feature selection, model building, validation, and model testing on an independent dataset Supervised learning classifiers can achieve highly accurate molecular classification if enough samples are available to "train" a classifier. In general, pairwise comparisons are less challenging than multi-class distinctions. In every case the comparison of a supervised classifier has to be done against the best generally accepted clinical classification method such as standard histopathology. In the next few sections we will review in more detail the steps necessary to select and validate gene markers and to build classifiers.

Selecting and Validating Gene Markers

Genes correlated with a binary class distinction, for example a morphological or clinical phenotype, can directly be identified and selected by using a "distance" metric, for example:

- Signal to noise ratio = $(\mu_A - \mu_B) / (\sigma_A + \sigma_B)$ [μ and σ are the means and std. dev. per class]
- t-test statistic = $(\mu_A - \mu_B) / \sqrt{}(\sigma^2{}_A + \sigma^2{}_B)$ [μ and σ are the means and std. dev. per class]
- Pearson correlation coefficient

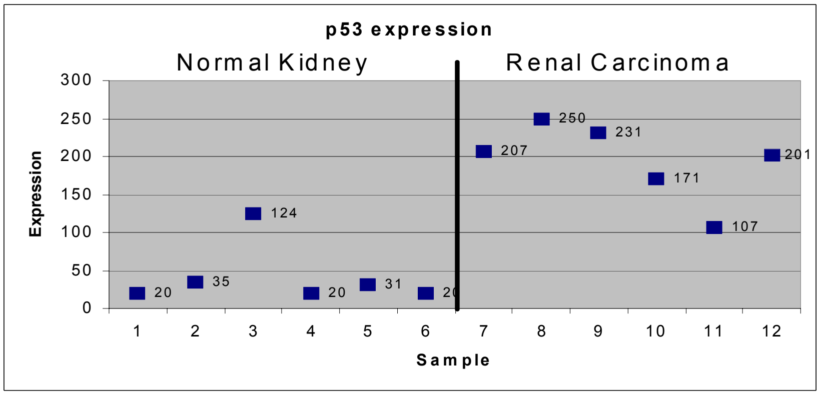
For example, the figure below shows the top 10 genes that differentiate normal kidney from renal carcinoma as selected from a microarray profiling experiment using the signal to noise (S2N) ratio score:



Number	S2N Score	Accession	Description
1	2.82	J03507	C7 Complement component 7
2	2.21	HG3431-HT3616	Decorin, Alt. Splice 1
3	2.08	Z30644	GB DEF = Chloride channel (putative)
4	2.07	J05257	DPEP1 Dipeptidase 1 (renal)
5	1.98	U27333	Alpha-1,3 fucosyltransferase 6 (FCT3A)
6	1.81	X56494	PKM2 Pyruvate kinase, muscle
7	1.78	X59798	CCND1 Cyclin D1
8	1.67	M22898	TP53 Tumor protein p53
9	1.51	D50855	CASR Calcium-sensing receptor
10	1.47	HG662-HT662	Epstein-Barr Virus Small Rna-Associated Protein

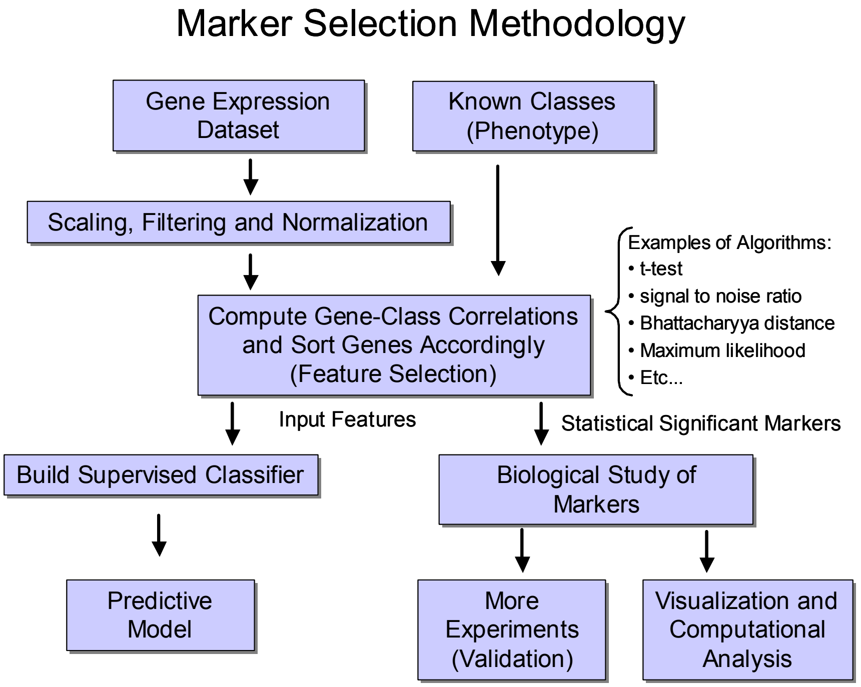
The original dataset was created by joining twelve microarray datasets, six from normal kidney and six from renal cell carcinoma samples. Markers were selected by computing the signal to noise score: the mean and standard deviation of the expression values are computed in each class and then the ratio of the difference of the means is divided by the sum of the standard deviations. For example, this calculation as applied to the profile of p53 shown below produces the following:

$$\text{Signal to noise ratio} = (\mu_{\text{cancer}} - \mu_{\text{normal}})/\ (\sigma_{\text{cancer}} + \sigma_{\text{normal}}) = 1.67$$



As can be seen, this gene acts as a marker of the "cancer" phenotype by being expressed on average at higher level in cancer samples compared with normal ones. It is important to notice that the difference in absolute expression value may not always be large. In this example p53 is a marker but in general displays low values of expression.

This basic procedure of selecting *differentially* expressed genes is useful in two common analysis situations. The first is associated with selecting statistically significant markers for more detailed follow up biological study (e.g. to identify genes that are differentially expressed in two different cancer types). Selected genes can then be subject to a literature search or to validation using other experimental assays (e.g. RT-PCR, immunohistochemistry, etc.). The second relates to the problem of *feature selection*: finding genes to feed into a supervised learning classifier. In this case one is interested in selecting the subset of genes most likely to be useful in discriminating phenotypes of interest, either as single markers or in combination with others. This task is better viewed as a pre-processing step in a classification methodology. Gene selection is required, in part, because many supervised learning algorithms perform sub-optimally with thousands of input variables and require some type of dimensionality reduction. A general methodology for supervised marker selection and classification is shown in figure 6. The training of classifiers will be discussed in detail in a subsequent section.



Marker Selection Methodology
Gene Expression Dataset
Known Classes (Phenotype)
Scaling, Filtering and Normalization
Examples of Algorithms:
• t-test
• signal to noise ratio
• Bhattacharyya distance
• Maximum likelihood
• Etc...
Compute Gene-Class Correlations and Sort Genes Accordingly (Feature Selection)
Input Features
Statistical Significant Markers
Build Supervised Classifier
Biological Study of Markers
Predictive Model
More Experiments (Validation)
Visualization and Computational Analysis

Figure 6. Methodology for marker selection.

Permutation Tests

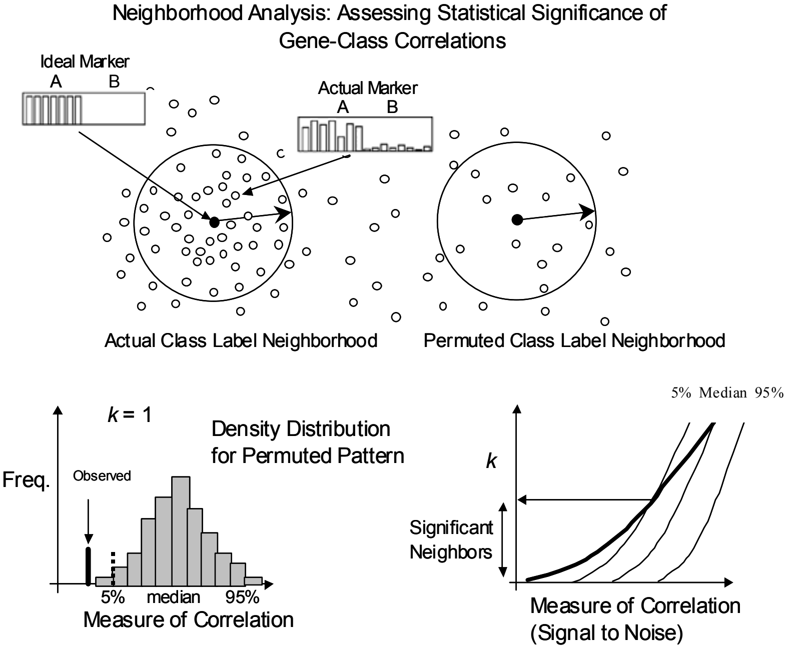
Once marker genes have been selected, one might want to decide how many of them to consider for further study. This is a difficult problem because typically there will be a gradual decrease in the score or correlations in such way that there is no well defined boundary between markers and non-markers. In most situations the analysis will concentrate on the very top markers and exclude the rest. However, this problem can be addressed more formally by using permutation testing. This method (Golub et al 1999, Slonim et al 2000) attempts to solve the marker selection problem by comparing the actual distribution of marker scores to a reference empirical distribution of scores obtained by permuting the phenotype class labels. The markers are viewed as close matches or “neighbors” of an ideal marker separating the classes. A histogram of scores for each of the ranked marker genes, corresponding to each permutation (neighborhood), is kept and the significance of an actual gene marker is obtained by finding the appropriate percentile in the histogram of the correspondingly ranked marker (i.e. the one with the same rank, e.g. best match, second best match etc.). There are several advantages to performing a permutation test: (1) the method does not assume a particular functional form for the distribution or correlation structure of genes; (2) it is performed on the entire distribution of marker genes and therefore takes into account the gene-to-gene correlation structure; and (3) it is a simple, intuitive approach that provides higher statistical power. In detail, the permutation test procedure for a given comparison of interest (e.g. markers high in class 0 and low in class 1) is as follows:

- Generate signal-to-noise $(\mu_{\text{class 0}} - \mu_{\text{class 1}})/(\sigma_{\text{ class 0}} + \sigma_{\text{class 1}})$ or other type of scores (t-test, Pearson etc.) for all genes being considered using the actual class labels (phenotype) and sort them accordingly. The best match (k=1) is the gene "closer" or more correlated to the phenotype using the signal to noise as a correlation function. In fact one can imagine the reciprocal of the signal to noise as a "distance" between the "phenotype" and each gene as shown in figure 7.

- Generate 500 or more random permutations of the class labels (phenotype). For each case of randomized class labels generate signal-to-noise scores and sort genes accordingly.

- Build a histogram of signal to noise scores for each value of k. For example one for all the 500 top markers (k=1), another one for the 500 second best (k=2) etc. These histograms represent a reference distribution for the *kth* marker and for a given value of k different genes contribute to it. Notice that the correlation structure of the data is preserved by this procedure. For each value of k, determine different percentiles (1%, 5%, 50% etc.) of the corresponding histogram.

- Compare the actual signal to noise scores with the different significance levels obtained for the histograms of permuted class labels for each value of k. This test helps to assess the statistical significance of gene markers in terms of the distribution of class-gene scores using permuted labels.



Neighborhood Analysis: Assessing Statistical Significance of Gene-Class Correlations
Ideal Marker
A B
Actual Marker
A B
Actual Class Label Neighborhood
Permuted Class Label Neighborhood
k = 1
Density Distribution for Permuted Pattern
Freq.
Observed
5% median 95%
Measure of Correlation
5% Median 95%
k
Significant Neighbors
Measure of Correlation (Signal to Noise)

Figure 7. Permutation test based assessment of significance for gene markers.

For example, normal kidney vs renal carcinoma marker selection and permutation testing for each of the selected markers generates the following list:

Number	Class	S2N Score	Perm 1%	Perm 5%	Median	Gene	Description
1	Normal	2.82	2.48	1.97	1.37	J03507	C7 Complement component 7
2	Normal	2.21	1.89	1.69	1.22	HG3431-HT3616	Decorin, Alt. Splice 1
3	Normal	2.08	1.83	1.56	1.12	Z30644	GB DEF = Chloride channel (putative)
4	Normal	2.07	1.76	1.47	1.07	J05257	DPEP1 Dipeptidase 1 (renal)
5	Normal	1.98	1.66	1.41	1.05	U27333	Alpha-1,3 fucosyltransferase 6 (FCT3A)
6	Carcinoma	1.81	2.38	1.97	1.39	X56494	PKM2 Pyruvate kinase, muscle
7	Carcinoma	1.78	1.99	1.74	1.21	X59798	CCND1 Cyclin D1
8	Carcinoma	1.67	1.82	1.58	1.13	M22898	TP53 Tumor protein p53 (Li-Fraumeni syndrome)
9	Carcinoma	1.51	1.72	1.48	1.07	D50855	CASR Calcium-sensing receptor
10	Carcinoma	1.47	1.66	1.43	1.04	HG662-HT662	Epstein-Barr Virus Small Rna-Associated Protein

The *class* column represents the class for which the markers are high (low in the other class). The *S2N score* is the signal to noise of each marker. The Perm 1%, 5% and 50% columns represent the percentiles in the histograms of signal to noise scores for permuted labels, for a given value of the rank order. These 10 markers shown all have signal to noise scores better than 5% of the random permutations (p <= 0.05).

Permutation tests assess the significance of gene markers in terms of class-gene correlations. If a group of genes fails to pass permutation testing, however, that by itself does not necessarily imply that it cannot be used to build an effective classifier (Huberty 1994, Kearns and Vazirani 1997). In subtle phenotypes distinctions, for example, the top marker genes are often “weak” and may not show overwhelming statistical significance. This often results from a gene being expressed only in a subset of samples in a given class. However, such genes can still be effective when used in combination as input to a classifier. Examples of this phenomenon can be found in subsequent sections.

Other marker selection methods have been introduced in the literature. For example the SAM method of Tusher et al 2001 is similar to the one presented above but includes a user-adjustable threshold to provide estimates of the false discovery rate. Dudoit et al 2001 have introduced a method based on step-down adjusted p-values using Westfall and Young’s approach in the context of replicated cDNA experiments. Ideker et al 2000 used generalized likelihood tests to assess the statistical significance of differentially expressed genes in the context of two channel cDNA microarrays. Newton et al 2001 and Baldi and Long 2001 used empirical Bayes hierarchical models to assess significance of differential expression. Lee et al 2000 combined the data from replicates to estimate posterior probabilities and identify differentially expressed genes. No systematic comparison of the error rates and statistical power of all these different methods have been published yet. Methods have also been proposed to combine both resampling and explicit control of the false discover rate (Yekuteli and Benjamini 1999) such as the stepwise permutation-based procedures of Korn et al 2002.

A logical extension of marker selection is *pattern discovery*, where one tries to find sub-patterns, i.e. patterns not necessarily involving all of the samples but that occur often and may represent groups of co-regulated or correlated genes. Califano et al (1999) introduced a pattern discovery algorithm (SPLASH) to expose more complex gene correlations. They extracted statistically significant subpatterns from expression array data using a geometric hashing algorithm. Although their statistical models were simplistic, their work represented one of the first analytic evaluations of sub-pattern significance in that context. Other attempts to elucidate complex gene-gene correlations or global correlation structure have used principal component analysis (PCA) (Bittner et al. 2000, Pomeroy et al 2002), singular value decomposition (Alter et al. 2000), biclustering (Cheng and Church 2000), and Plaid (Lazzeroni and Owen 2000). Hastie and associates introduced “gene shaving” as a global approach based on PCA to systematically expose coherent patterns of co-regulation in gene expression data (Hastie et al. 2000). All these methods are promising but face the same challenge in terms of how to effectively separate biologically relevant signals from the noise.

Class Prediction

A general methodology for class prediction under the supervised learning paradigm is shown in figure 11. One starts by putting together the relevant samples into a single dataset, scaling and pre-processing the dataset and by defining the target phenotype class based on morphology, tumor type or treatment outcome clinical information. The dataset is split in train and test subsets if enough samples are available. If not enough samples are available, one can perform a leave-one out cross validation in which one samples is held, a predictor is trained on the remaining samples, the left out sample is classified by this predictor, and the process is repeated iteratively. Once a proper training set has been defined, a marker selection methodology is applied. This step is in general useful and facilitates the training of most classification algorithms, although some classifiers such as Naïve Bayes or Support Vector Machines can deal with thousands of variables effectively (Ramaswamy 2001a, Weston et al 2001). Feature selection is generally useful to facilitate subsequent validation of selected genes that are particularly informative in classification. Once markers have been selected, a classifier can be built using classification algorithms such as (Duda et al 2000, Fukunaga 1990, Ripley 1996):

- Linear or Quadratic Discriminants
- *k*-Nearest Neighbors
- Weighted Voting
- Naïve Bayes
- Neural Networks
- Support Vector Machines
- Decision Trees

If the model has internal parameters that require tuning, this is done typically when training the predictor. In this way several models are built using different number of marker genes and the final chosen model is the one that minimizes the total error in cross-validation. This model can then be validated on an independent test set. Detailed model-to-model performance comparisons require predictions with different instantiations of the train and test datasets and have to be made carefully as suggested by Salzberg (1999).

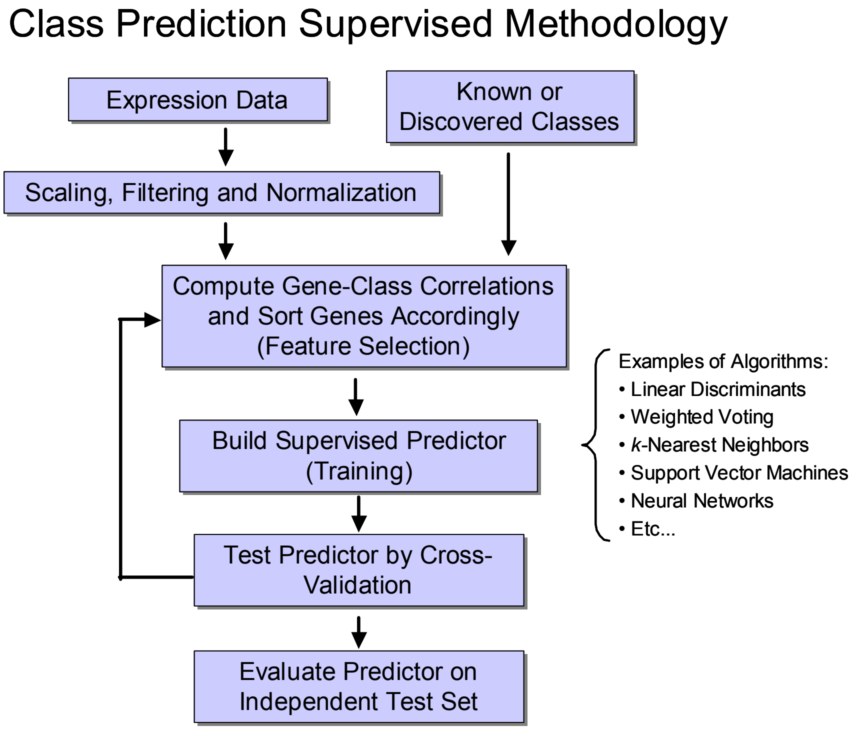
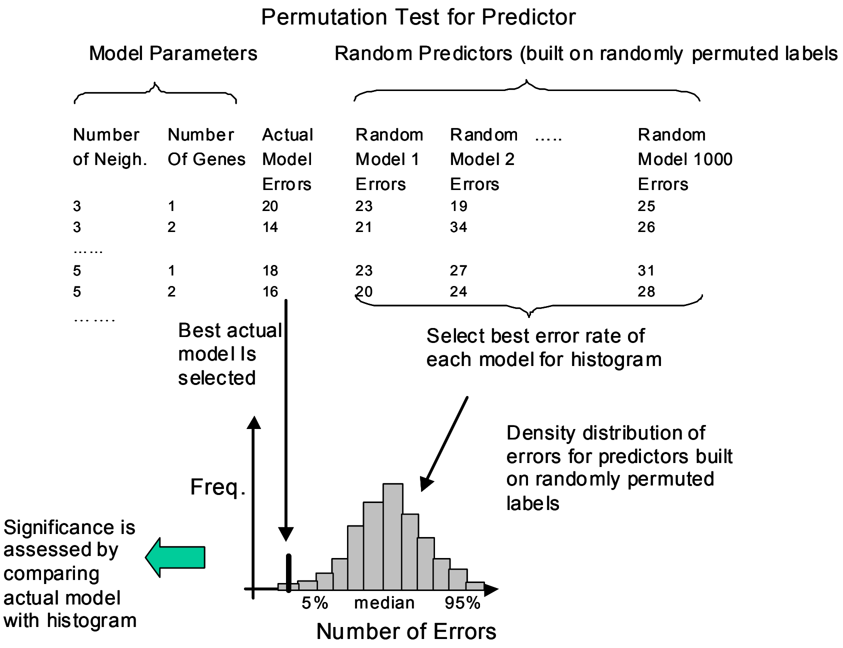

Class Prediction Supervised Methodology
Expression Data
Known or Discovered Classes
Scaling, Filtering and Normalization
Compute Gene-Class Correlations and Sort Genes Accordingly (Feature Selection)
Build Supervised Predictor (Training)
Test Predictor by Cross-Validation
Evaluate Predictor on Independent Test Set
Examples of Algorithms:
• Linear Discriminants
• Weighted Voting
• k-Nearest Neighbors
• Support Vector Machines
• Neural Networks
• Etc...

Figure 11. Methodology for Supervised Learning.

Statistical Significance of a Supervised Classifier.

The statistical significance of a supervised classifier can be evaluated in several ways. One of the simplest is to compute a Fisher exact test of the classification confusion matrix or use the proportional chance criterion to compare the observed with the expected classification accuracy for a random predictor (Huberty 1994). A more sophisticated empirical approach, sometimes useful for weak classifiers or when there are not enough samples to create an independent test set and when cross-validation must be used, is the class label permutation (Fisher 1935, Lehman 1986, Good 1994). The phenotype (sample) labels are randomly permuted 1000 or more times and in each instance predictive models are built and tested. Once this is done one selects the best error rate for each of these 1000 random predictors and makes a histogram of these error rates. The error rate from the actual predictive model is then compared to this histogram to determine the statistical significance of this prediction (see figure 12).


Permutation Test for Predictor
Model Parameters
Random Predictors (built on randomly permuted labels
Best actual model Is selected
Select best error rate of each model for histogram
Density distribution of errors for predictors built on randomly permuted labels
Freq.
Significance is assessed by comparing actual model with histogram
5%
median
95%
Number of Errors

Number of Neigh.	Number Of Genes	Actual Model Errors	Random Model 1 Errors	Random Model 2 Errors	…..	Random Model 1000 Errors
3	1	20	23	19		25
3	2	14	21	34		26
……						
5	1	18	23	27		31
5	2	16	20	24		28
…….						

Figure 12. Methodology to assess the statistical significance of a classifier.

Figure 13 below shows the application of this permutation test for the k-nearest neighbor treatment outcome predictor in Pomeroy et al 2002. This is a cross-validation model built on 60 medullobalstoma samples capable of distinguishing patients with “good” and “poor” prognosis on the basis of primary tumor gene expression profiles. An optimal model was defined using the following parameters:

Number of neighbors (k): 3, 5
Number of genes (ng): 1,2,3,4,5,6,7,8,9,10,15,25,50,100

Models were created using the actual treatment outcome labels and also for 1000 random permutation of those labels (keeping the gene expression data the same). The best predictive model used k=5 and ng=8, and correctly predicted 47 out of 60 cases as being “good” or “poor” prognosis. Random class label permutation showed that there were 9 models with better performance (lower error rates) than the actual model. Based on this result, the statistical significance of this medulloblastoma outcome prediction study was p = 0.009 (9/1000). When enough samples are available to produce independent train and test datasets the proportional chance criterion is usually a sufficient measure of statistical significance (Huberty 1994).

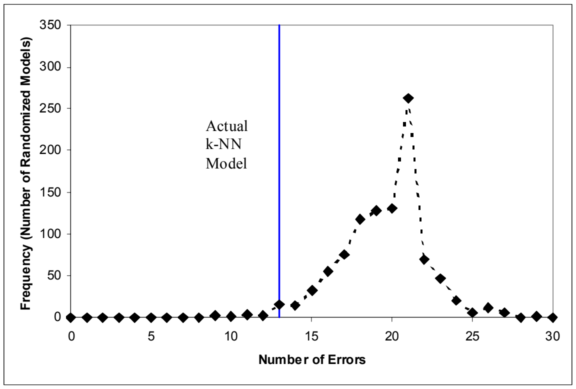


Figure 13. Results of the permutation test for a k-nearest neighbor Medulloblastoma treatment outcome predictor.

Pairwise Classification: Classifying Leukemia Subtypes

We next review microarray-based leukemia sub-classification (Golub et al 1999, Slonim et al 2000) as an example of a binary molecular classification problem. Acute leukemias arise from different precursor cells: lymphoid (acute lymphoblastic leukemia (ALL)) and myeloid (acute myeloid leukemia (AML)). This distinction is critical for effective Leukemia treatment planning, and is currently done by assimilation of diverse information including morphological, cytogenetic, histochemical, and immunophenotypic analysis by an expert physician. Our initial analysis employed a set of 27 ALL and 13 AML samples. A permutation test of the gene markers revealed a striking excess density of genes correlated with the class distinction. We decided to employ a weighted voting classifier based on the top 50 genes. Sets of classifiers were first constructed in cross-validation experiments using the 40 leukemia samples. In one case, no prediction was made because the confidence score fell below a predetermined threshold. For the remaining 39 cases, the prediction accuracy was 100%. While they initially chose 50 genes for the prediction algorithm, they also found that classifiers involving as few as 7 genes proved to be 100% accurate in the ALL/AML distinction. Interestingly, however, among the top 50 genes, no single gene yielded a perfect predictor. Correct classification thus requires multi-gene predictors. Other classification algorithms such as Naïve Bayes, k-nearest neighbors and Support Vector Machines produce similar results (Mukherjee, et al 1999).

The original weighted voting classifier was also tested on an independent collection of 34 AML and ALL samples. In 3 cases, the confidence score fell below the threshold for prediction but the classifier made predictions in the remaining 30 cases, and 29 out of 30 were correct. The single error had the lowest confidence score of the samples, just barely passing the threshold. Overall, 69 of 70 samples were correctly classified either in cross-validation or using the independent test set (98.6%). Other algorithms also performed fairly well on the independent test set with the Support Vector Machine model producing 100% accuracy.

The marker genes, shown in figure 14 below, are highly instructive. Some, including CD22, CD11c, CD33 and CD79a, encode cell surface proteins for which monoclonal antibodies have been previously demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells. Others provide new markers of acute leukemia subtype. For example, the leptin receptor, originally identified as a cell surface receptor in adipocytes, but showed high relative expression in AML cells. The leptin receptor has been demonstrated to have anti-apoptotic function in hematopoietic cells. Some of the markers are typical markers of hematopoetic lineage but others have biological function relevant to the cancer. For example many of the genes encode proteins critical for S-phase cell cycle progression (Cyclin D3, Op18 and MCM3), chromatin remodeling (RbAp48), transcription (SNF2b and TFIIEß), or cell adhesion (zyxin and integrin alpha X) or are known oncogenes (c-MYB, E2A, EWSR1 and HOXA9).

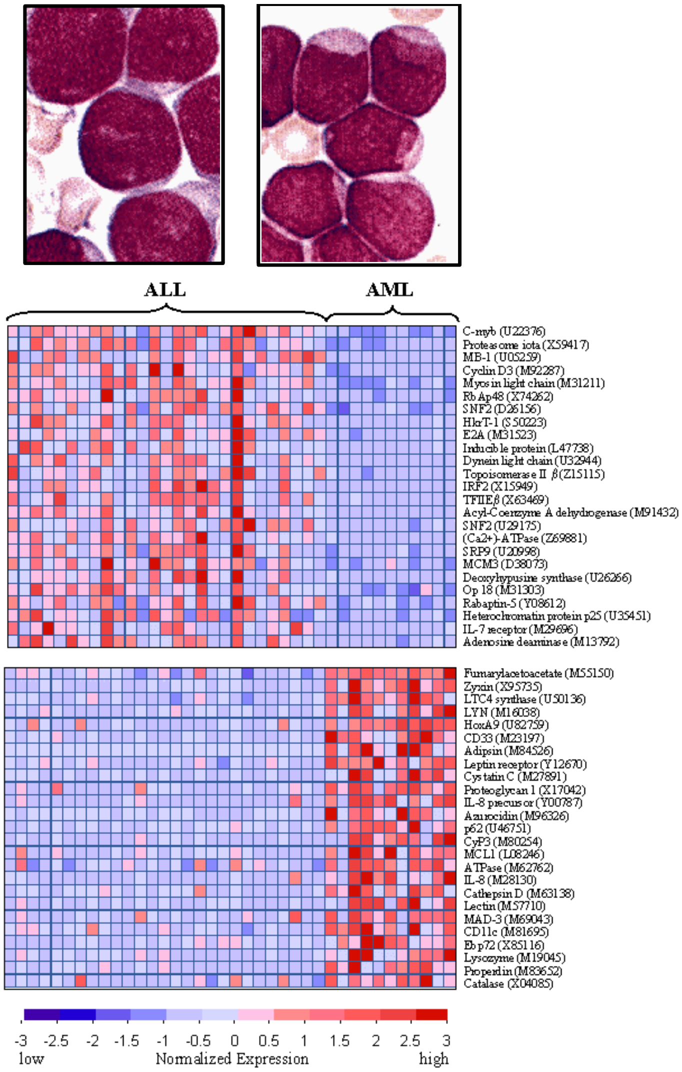


Figure 14. Top markers of the ALL/AML leukemia subtype distinction. The micrographs on top show the similar morphology characteristic of these cells.

Predicting Treatment Outcome: Lymphoma

Supervised learning classifiers are also well suited to predict differential treatment outcome between histologically similar tumors. Here we review the results of Lymphoma treatment outcome prediction model of Shipp et al 2001. Diffuse Large B-Cell Lymphomas (DLBCL) are the most common lymphoid neoplasm and it accounts for up to 40% of adult (non-Hodgkin's) lymphomas. Using existing chemotherapeutic regimens only a subset of DLBCL patients is cured. Clinical prognostic models such as the International Prognostic Index (IPI) are used to identify different DLBCL risks groups. The clinical factors used by the IPI (age, performance status, stage, number of extranodal sites, and serum LDH) are potentially surrogate markers for the true molecular heterogeneity of the disease and provide a useful but highly imperfect model for the identification of high-risk patients. Few molecular markers are, however, broadly useful for lymphoma risk stratification.

Our group studied 58 DLBCL patients uniformly treated with standard CHOP chemotherapy, where long-term clinical follow-up was available (Shipp et al (2001)). These patients fell into two groups including those with cured disease and those with fatal / refractory disease. They used supervised learning to determine differential treatment outcome on the basis of primary tumor gene expression profiles.

Top marker genes for the cured vs. failure distinction were selected using the signal to noise ratio (figure 17).

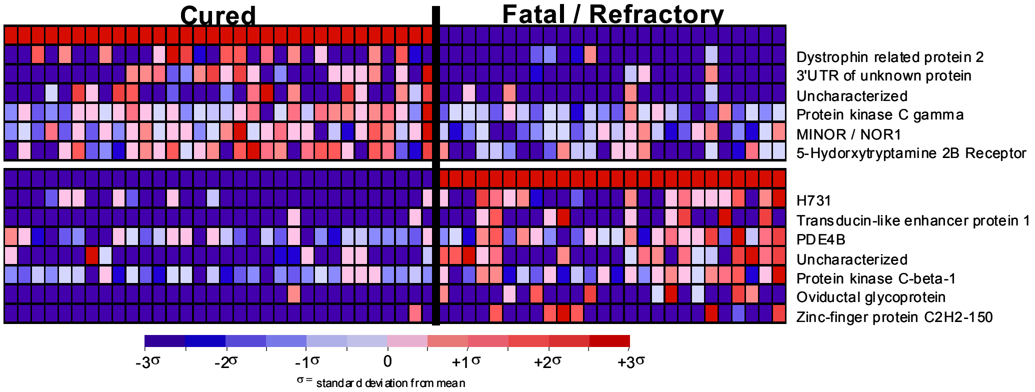


Figure 17. Top markers of Lymphoma treatment outcome.

We developed a supervised classifier using a Weighted Voting algorithm (Slonim et al 2000) and used cross-validation testing to assess the performance of the classifier. Models containing between 8 and 16 genes yielded statistically significant predictions with the highest accuracy obtained using 13 genes. This classifier separated the 58 patients into 2 groups according to the predicted class: predicted to be cured or predicted to have fatal / refractory disease based on the gene expression profiles of those 13 genes. A Kaplan-Meier plot of these results is shown in Figure 18 (p = 0.0013 using a standard log-rank test).

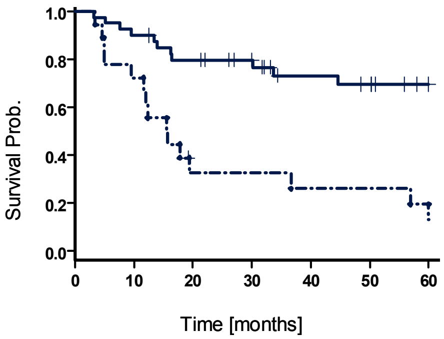


Figure 18. Kaplan-Meier survival plot for the treatment outcome predicted groups.

Patients predicted by the classifier to be cured had dramatically improved long-term survival compared to those predicted to have fatal/refractory disease. The 5-year OS is 70% vs. 12%, with nominal log rank P-value of 0.00004. As part of this study, we also built other supervised classification algorithms and obtained similar results. The fact that treatment outcome can be predicted solely based on gene expression patterns indicates the existence, at diagnosis, of a gene expression signature of outcome in DLBCL.

Multi-Class Classification: Classifying Multiple Tumor Types

Multiclass classification problems are inherently more difficult than pairwise comparisons. In this section we review our efforts to perform multiclass tumor classification (Ramaswamy et al 2001, Yeang et al 2001). We explored the general feasibility of molecular cancer diagnosis of common human tumors solely on the basis of tumor gene expression profiles. We first created a gene expression database containing the expression profiles of 218 tumor samples representing 14 common human cancer classes and devised a multiclass classification method. Our analytical scheme is depicted in figure 15. First, The multiclass problem was divided into a series of 14 one-versus-all (OVA) pairwise comparisons. Each test sample was presented sequentially to these 14 pairwise classifiers, each of which either claimed or rejected that sample as belonging to a single class. This method resulted in 14 separate OVA classifications per sample, each with an associated confidence. Each test sample was then assigned to the class with the highest OVA classifier confidence. In mathematical terms: given *m* classes and *m* trained classifiers, a new sample takes the class of the classifier with the largest real valued output *class* = *arg* $max_{i=1...m}$ f_i, where f_i is the real valued output of the *i*th classifier. A positive prediction strength corresponds to a test sample being assigned to a single class rather than to the "all other" class.

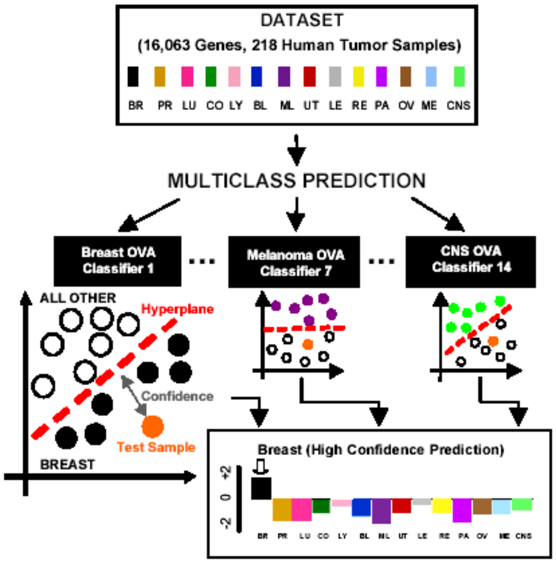

DATASET
(16,063 Genes, 218 Human Tumor Samples)
BR PR LU CO LY BL ML UT LE RE PA OV ME CNS
MULTICLASS PREDICTION
Breast OVA Classifier 1
...
Melanoma OVA Classifier 7
...
CNS OVA Classifier 14
ALL OTHER
Hyperplane
Confidence
Test Sample
BREAST
Breast (High Confidence Prediction)
-2 0 +2
BR PR LU CO LY BL ML UT LE RE PA OV ME CNS

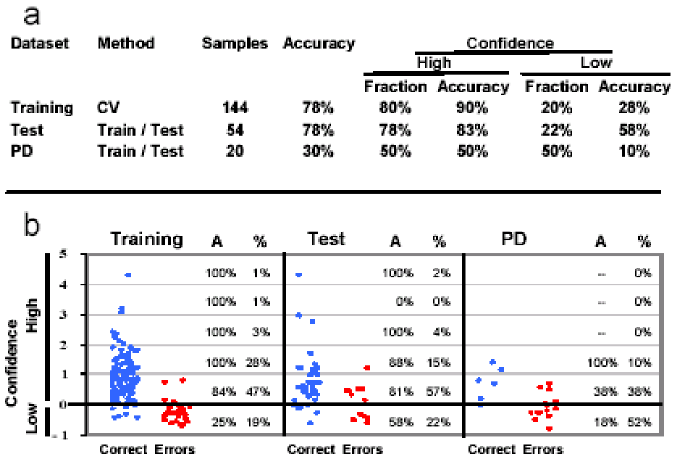
Figure 15. Multiclass classification scheme. The multiclass cancer classification problem is divided into a series of 14 OVA problems, and each OVA problem is addressed by a different class-specific classifier (e.g., "breast cancer" vs. "not breast cancer"). Each classifier uses the SVM algorithm to define a hyperplane that best separates training samples into two classes. In the example shown, a test sample is sequentially presented to each of 14 OVA classifiers and is predicted to be breast cancer, based on the breast OVA classifier having the highest confidence.

We then evaluated several classification algorithms for these OVA pairwise classifiers including weighted voting, k-nearest neighbors, and Support Vector Machines (SVM). Because the SVM algorithm consistently out-performed other algorithms, these results are described in detail. The

SVM algorithm was used recently for pairwise gene expression-based classification (Mukherjee et al 1999, Brown et al 2000, Weston et al 2001) and has a strong theoretical foundation (Vapnik 1996, Evgeniou et al 2000). This algorithm considers all profiled genes, to create descriptions of samples in this high-dimensional space, and then defines a hyperplane that best separates samples from two classes (Figure 15). The position of an unknown sample relative to the hyperplane determines its membership in one or the other class (e.g., ‘‘breast cancer’’ vs. ‘‘not breast cancer’’). Fourteen separate SVM-based OVA classifiers classify each sample. The confidence of each OVA SVM prediction is based on the distance of the test sample to each hyperplane, with a value of 0 indicating that a sample falls directly on a hyperplane. The overall multiclass classifier assigns a sample to the class with the highest confidence among the 14 pairwise OVA analyses.

The accuracy of this multiclass SVM-based classifier in cancer diagnosis was first evaluated by cross-validation in a set of 144 training samples. This method involves randomly withholding 1 of the 144 primary tumor samples, building a predictor based only on the remaining samples, then predicting the class of the withheld sample. The process is repeated for each sample, and the cumulative error rate is calculated. As shown in Figure 16, the majority (80%) of the 144 calls were high confidence (defined as confidence >= 0) and these had an accuracy of 90%, using the patient’s clinical diagnosis as the ‘‘gold standard.’’ The remaining 20% of the tumors had low confidence calls (confidence < 0), and these predictions had an accuracy of 28%. Overall, the multiclass prediction corresponded to the correct assignment for 78% of the tumors. For half of the errors, the correct classification corresponded to the second- or third-most confident OVA prediction.

These results were confirmed by training the multiclass SVM classifier on the entire set of 144 samples and applying this classifier without further modification to an independent test set of 54 tumor samples. Overall prediction accuracy on this test set was 78%, a result similar to cross-validation accuracy and highly statistically significant when compared with class-proportional random prediction *(P* < 10^{16}). The majority of these 54 predictions (78%) were high confidence, with an accuracy of 83%, whereas low-confidence calls were made on the remaining 22% of tumors with an accuracy of 58%. Again for one-half of the errors, the correct classification corresponded to the second-or third-best prediction. Of note, classification of 100 random splits of a combined training and test dataset gave similar results, confirming the stability of prediction for this collection of samples.



a

Dataset	Method	Samples	Accuracy	Confidence			
				High		Low	
				Fraction	Accuracy	Fraction	Accuracy
Training	CV	144	78%	80%	90%	20%	28%
Test	Train / Test	54	78%	78%	83%	22%	58%
PD	Train / Test	20	30%	50%	50%	50%	10%

Figure 16. Multiclass classification results. (a) Results of multiclass classification by using cross-validation on a training set (144 primary tumors) and independent testing with 2 test sets: Test (54 tumors; 46 primary and 8 metastatic) and PD (20 poorly differentiated tumors; 14 primary and 6 metastatic). (b) Scatter plot showing SVM/OVA classifier confidence as a function of correct calls (blue) or errors (red) for Training, Test, and PD samples. A, accuracy of prediction; %, percentage of total sample number.

We next focused on the 28 samples that yielded low-confidence predictions in cross-validation, as the multiclass predictor generally misclassifies these samples. We found that a large number (17 of 28) were moderately or poorly differentiated (high-grade) carcinomas. It can be difficult to classify such tumors with traditional methods because they often lack the characteristic morphological hallmarks of the organ from which they arise. It has been assumed that these tumors are nonetheless fundamentally molecularly similar to their better-differentiated counterparts, apart from a few differences that might account for their clinically aggressive nature. To directly test this hypothesis, the multiclass classifier was trained on the original 144-tumor dataset, and then applied to an independent set of poorly differentiated tumors. Gene expression data were collected from 20 poorly differentiated adenocarcinomas (14 primary and 6 metastatic), representing 5 tumor types: breast, lung, colon, ovary, and uterus. The technical quality of this dataset was indistinguishable from the other samples in the study. However, these tumors could not be accurately classified according to their tissues of origin, compared with the high overall accuracy seen with lower-grade tumors. Overall, only 6 / 20 samples (30%) were correctly classified, which is statistically no better than what one would expect by chance alone *(P* = 0.38). Because the classifier relies on the expression of thousands of similarly weighted tissue-specific molecular markers to determine the class of a tumor, these findings indicate that poorly differentiated tumors do not simply lack a few key markers of differentiation, but rather have fundamentally distinct gene expression patterns.

Dimensionality Reduction and Projection: Principal Components Analysis

Datasets with a large number of genes are in general difficult to visualize. Principal Component Analysis (PCA) is a dimensionality reduction method that has been used to visualize complex gene expression datasets in two and three-dimensional plots (Mardia et al 1979, Yeung and Ruzzo 2001, Bittner et al. 2000, Pomeroy et al 2002). In this approach one finds standardized linear combinations of variables, the “principal components,’ which are orthogonal and explain all of the variance in the original dataset. A typical method to obtain a simple projection (multi-dimensional scaling) of the dataset is to plot the top 2 or 3 principal components, which may account for a significant fraction of the variance, in a scatter plot. One can take this approach in a completely unsupervised manner, e.g. by using all genes that pass a data “pre-processing” step, or in a supervised way by projecting only the top marker genes of a phenotype of interest.

For example, principle component analysis can be applied to leukemia gene expression data. The initial set of genes is first subject to a variation filter resulting in a dataset with 612 genes that displayed the greatest variation across samples. In this case the PCA method is used in an unsupervised way. Figure 19 shows a 3D plot of these leukemia samples projected in the space of the top 3 principal components. This plot reveals the dominant structure of the dataset corresponding to the known morphological subclasses of leukemia, clearly separating ALL from the AML samples and separating the T-ALL from B-ALL samples.

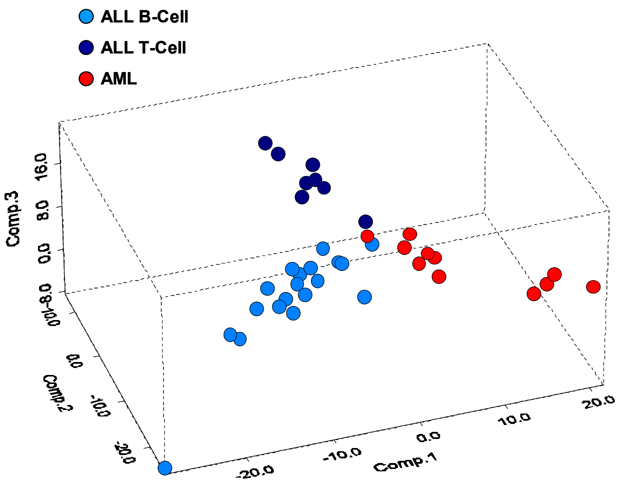


Figure 19. Plot of the top 3 principal components for the 612 most highly varying genes in the Leukemia subtypes dataset. The analysis is unsupervised and reveals the dominant structure of the dataset corresponding to the morphological subclasses.

Conclusions and Analytical Challenges in Molecular Classification

The analysis of cancer gene expression data is still in its infancy despite impressive recent progress. As expression profiling technologies mature, the identification of statistically significant patterns from relatively sparse and noisy data sets remains a major challenge. Although sophisticated data-mining techniques are already being used to analyze expression data, most of these techniques achieve robust performance with a large number of samples and a small number of variables (Friedman 1994). However, gene expression data sets generally contain small numbers of samples, many profiled genes, and multiple sources of variation. Future advances will require adapting analytic and statistical techniques to this type of data. In addition, most published work has analyzed a relatively small number of samples and most studies await independent confirmation.

A first generation of gene expression analysis methods has been used successfully in a variety of clustering and classification settings. For example, relatively successful models have been used to classify a variety of cancer types. Some examples include:

- Leukemias (Golub et al 1999, Yeoh et al 2002, Armstrong et al 2001).
- Lymphomas (Alizadeh et al 1999, Alizadeh et al 2000, Shipp et al 2001, Li et al 2002).
- Ewing's Sarcoma (Lessnick et al 2001).
- Brain cancer (Pomeroy et al 2002).
- Breast cancer (Porou et al 1999, Perou et al 2000, Sorlie et al 2001).
- Lung cancer (Bhattacharjee et al 2001, Garber et al 2001).
- Prostate cancer (Singh et al 2002, Welsh et al 2001a).
- Colon cancer (Alon et al 1999).
- Gastrointestinal tumors (Allander et al 2001).
- Ovarian cancer (Welsh et al 2001b).
- Melanoma (Bittner et al 2000).
- Multiple tumors (Ramaswamy et al 2001a, Su et al 2001).
- Soft tissue tumors (Nielsen et al 2002).

These studies have undoubtedly contributed to improve our understanding of cancer classification at the molecular level. However, in most cases the complexity of the problem had to be simplified by treating genes as independent variables. While some studies expose co-regulation, they may not focus on the more complex patterns of interaction inherent in all biological processes and may further ignore the diversity of biological mechanisms within a phenotype. For example, in marker selection, one distinguishes between two phenotypes by determining which genes are up-regulated in one phenotype and down-regulated in the other. While this is a straightforward pattern to discover, we know it does not represent the true nature of genes' interactions. For example, it does not take into account 1) distinct mechanisms that may yield the same biological state, or 2) sub-phenotypes and taxonomies that may be as yet unidentified. Even when clustering and classification methods are shown to be successful, it is often unclear exactly what the significant features or discovered patterns mean. Extracting more refined knowledge from the profiles and patterns is a serious scientific bottleneck.

Another important area relates to the integration of datasets generated in different laboratories using different profiling technologies. Many human cancer studies involve valuable or rare clinical specimens and are difficult to repeat. Ideally, one should be able to compare expression data sets obtained in any center, at any time, using any platform. However, this goal remains unrealized. Spotted array data is usually reported as ratios between experimental and control expression values and cannot be easily compared with oligonucleotide microarray data. Multiple expression profiling technologies require more sophisticated methods for data comparison and integration.

Despite initial sequencing of the human genome, we still have only a rudimentary knowledge of the physiologic roles of most genes. This represents a significant bottleneck in linking gene expression profiles to molecular mechanisms of transformation. There is a need for integrated databases, with complete annotation, comprehensive gene descriptions, and links to relevant genetic and proteomic information. In addition, as expression studies are performed in various species, integration of this information should prove as illuminating as inter-species gene sequence comparisons. Such databases will allow for an understanding of gene expression in the context of all other available biologic information. Although a number of commercial sources have started to create such databases, there is much room for improvement.

The challenges described above concern methodological and scientific issues. However, no computational approach is useful if it is not embodied in a set of software tools that scientists in the community can use. There are some academic codes available by web download, but often they are not integrated and do not interoperate in a user-friendly environment. Available commercial codes are generally not current with the latest sophisticated techniques and often focus more on visualization of expression data than analysis and knowledge discovery. Since analysis of gene expression data remains a significant limitation in cancer genomics, the development of freely available and transparent analytic software continues to remain a major challenge.

Acknowledgements

We are indebted to members of the Cancer Genomics Group, Whitehead / MIT Center for Genome Research and the Golub Laboratory, Dana-Farber Cancer Institute / Harvard Medical School for many valuable and stimulating discussions.

References

Alizadeh A, Eisen M, Davis RE, Ma C, Sabet H, Tran T, Powell JI, Yang L, Marti GE, Moore DT, Hudson Jr. JR, Chan WC, Greiner T, Weisenburger D, Armitage JO, Lossos I, Levy R, Botstein D, Brown PO, Staudt LM. (1999). The Lymphochip: A specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. Cold Spring Harbor Symposia on Quantitative Biology. 1999; 64(): 71-78.

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., Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M., et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511.

Allander, Susanne V., Nina N. Nupponen, Markus Ringne´r, Galen Hostetter, Greg W. Maher, Natalie Goldberger, Yidong Chen, John Carpten, Abdel G. Elkahloun, and Paul S. Meltzer (2001). Gastrointestinal Stromal Tumors with KIT Mutations Exhibit a Remarkably Homogeneous Gene Expression Profile. Cancer Research 61, 8624–8628, December 15, 2001.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA, Vol. 96, Issue 12, 6745-6750, June 8, 1999.

Alter, O., Brown, P.O., Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97:10101–10106.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J. (2001) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics 30, pp 41 - 47 (2002)

Baldi, P. Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509–519.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., Meyerson, M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* 98:13790–13795.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406:536–540.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr., Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97:262–267.

Califano, A., Stolovitzky, G., Tu, Y. (1999) Analysis of gene expression microarrays for phenotype classification. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, San Diego, August 19–23, pp. 75–85.

Cheng Li and Wing Hung Wong (2001a) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, Genome Biology 2(8): research0032.1-0032.11.

Cheng Li and Wing Hung Wong (2001b) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proc. Natl. Acad. Sci. Vol. 98, 31-36.

Cheng, Y., Church, G.M. (2000) Biclustering of expression data. *Proceedings of Intelligent Systems in Molecular Biology 2000*, August 19–23, 2000, La Jolla, CA.

Chiang et al. 2001. Compute genome-mean expression profiles from expression and sequence data.. Bioinformatics 17(S1), 49-55, 2001.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2:65–73.

Duda, R.O., Hart, P.E., Stork, D.G. (2000) *Pattern Classification*, 2ed. John Wiley & Sons, Inc., New York, NY.

Dudoit, S., Yang, Y. H, Speed, T.P., Callow, M.J. (2001) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* (in press).

Eisen, M., Spellman, P., Brown, P., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–14868.

Evgeniou T., Pontil, M., Poggio, T. (2000) Regularization networks and support vector machines. *Advances in Computational Mathematics* 13:1–50.

Fisher, R. The Design of Experiments. 3ed. Oliver and Boyd Ltd. London. 1935.

Friedman, J.H. (1994) An overview of computational learning and function approximation. In *From Statistics to Neural Networks. Theory and Pattern Recognition Applications*, Cherkassy, V., Friedman, J., Wechsler, H.W., eds. Springer-Verlag, New York.

Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition*, 2ed. Academic Press, New York.

Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A. 2001 Nov 20; 98(24): 13784-9.

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., Lander, E.S. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.

Good, P. (1994) Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer-Verlag, New York.

Hastie, T., Tibshirani, R., Botstein, D., Brown, P. (2001) Supervised harvesting of expression trees. *Genome Biol.* 2:1–12.

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown, P. (2000) "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 1:Research0003.1–0003.21.

Huberty, C. J."Applied Discriminant Analysis," John Wiley and Sons Inc. (1994).

Hunter, L.; Taylor, R. C.; Leach, S. M.; and Simon, R., GEST: A Gene Expression Search Tool Based on a Novel Bayesian Similarity Metric, Bioinformatics 17: 115S-122S (2001).

Ideker, T., Thorsson, V., Siegel, A.F., Hood, L.E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.* 7:805–817.

Kearns M. J. and U. V. Vazirani, "An Introduction to Computational Learning Theory", MIT Press. 1997.

Kerr A. and G. Churchill (2001a), Experimental design for gene expression microarrays, Biostatistics, 2:183-201 (2001).

Kerr A. and G. Churchill (2001b), Statistical design and the analysis of gene expression microarrays, Genetical Research, 77:123-128, 2001.

Kerr, Afshari, Bennett, Bushel, Martinez, Walker and Churchill (2001), Statistical analysis of a gene expression microarray experiment with replication, Statistica Sinica, to appear.

Korn, Edward, James F. Troendle, Lisa M. McShane, and Richard Simon. Controlling the number of false discoveries. Application to high-dimensional genomic data. NCI –DCTD-003Technical report (http://linus.nci.nih.gov/~brb/TechReport.htm)

Lazzeroni, L., Owen, A.B. (2000) Plaid models for gene expression data. http://www-stat.stanford.edu/~owen/reports/plaid.pdf.

Lee, M.T., Kuo, F., Whitmore, G.A., Sklar, J. (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* 97:9834–9839.

Lehman, E.C. (1986) *Testing Statistical Hypothesis*, 2ed. John Wiley & Sons, Inc., New York, NY.

Lessnick, Stephen L, Caroline S. Dacwag, and Todd R. Golub. The Ewing's Sarcoma Oncoprotein EWS/FLI Induces a p53-Dependent Growth Arrest in Primary Human Fibroblasts. Cancer Cell 1, 393-401, 2001.

Li S, Ross DT, Kadin ME, Brown PO, Wasik MA. Comparative genome-scale analysis of gene expression profiles in T cell lymphoma cells during malignant progression using a complementary DNA microarray. Am J Pathol. 2001 Apr; 158(4): 1231-7.

Mardia, K.V., Kent, J.T., Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press, London.

Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J.P., Poggio, T. (1999) Support vector machine classification of microarray data. CBCL Paper #182 Artificial Intelligence Lab. Memo #1676, Massachusetts Institute of Technology, Cambridge, MA, December 1999.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., Tsui, K.W. (2001) On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* 8:37–52.

Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, van de Rijn M. Molecular characterisation of soft tissue tumours: a gene expression study. Lancet. 2002 Apr 13; 359(9314): 1301-7.

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., Lashkari, D., Shalon, D., Brown, P.O., Botstein, D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* 96:9212–9217.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A.L., Brown, P.O., Botstein, D. (2000) Molecular portraits of human breast tumours. *Nature* 406:747–752.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E., Golub, T. (2001) Gene expression-based classification and outcome prediction of central nervous system embryonal tumors. Nature, Vol 415, 24 (2002).

Ramaswamy S, Osteen RT, Shulman LN (2001b). Metastatic Cancer from an Unknown Primary Site, in Lenhard RE, Osteen RT, Gansler T (eds): Clinical Oncology (ed1). Atlanta, GA, American Cancer Society, 2001, pp.711-719.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R. (2001a) Multiclass cancer diagnosis by using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* 98: 15149-15154.

Ripley B. D.. Pattern Recognition and Neural Networks, Cambridge: University Press (1996).

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet. 2000 Mar; 24(3): 227-35.

Salzberg S. On comparing classifiers: A critique on current research and methods. Data Mining and Knowledge Discovery 1, 1-12, 1999.

Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN. A gene expression database for the molecular pharmacology of cancer.

Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M.,Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, T., Mesirov, J., Neuberg, D., Lander, E., Aster, J.,Golub, T. (2001) Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nature Medicine January 2002 Volume 8 Number 1 pp 68 – 74.

Singh, Dinesh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, William R. Sellers. Gene Expression Correlates of Clinical Prostate Cancer Behavior. Cancer Cell: March 2002, Vol. 1.

Slonim, D.K., Tamayo, P., Mesirov, J.P., Golub, T.R., Lander, E.S. (2000) Class prediction and discovery using gene expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB) 2000*. ACM Press, New York, pp. 263–272.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001 Sep 11; 98(19): 10869-74.

Su, A., Welsh, J.B., Sapinoso, L.M., Kern, S., Lapp, H. Dimitrov, P., Schultz, P.G., Powell, S., Moskaluck, C., Frierson, H.F. Jr. and Hampton, G.M. (2001). Molecular classification of human carcinomas using gene expression signatures. Cancer Research, 61, 7388-7393.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Dmitrovsky, E., Lander, E.S., Golub, T.R. (1999) Interpreting gene expression with self-organizing maps: Methods and application to hematopoeitic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907–2912.

The Chipping Forecast (1999) Special Supplement. *Nature Genet.* 21:January 1999.

Tibshirani, Robert, Guenther Walther and Trevor Hastie. Estimating the number of clusters in a dataset via the Gap statistic" JRSSB 2000.

Tseng George, Min-Kyu Oh, Lars Rohlin, James C Liao and Wing Hung Wong (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variation and assessment of gene effects. Nucleic Acids Research, Vol. 29, No. 12. 2549-2557.

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

Vapnik, V.N. (1998) *Statistical Learning Theory*, John Wiley & Sons, Inc., New York.

Welsh, J.B., Sapinoso, L.M., Kern, S., Wang-Rodriguez, J., Moskaluk, C., Frierson, H.F. Jr. and Hampton, G.M. (2001a) Analysis of gene expression identifies candidate molecular markers and pharmacologic targets in prostate cancer. Cancer Research, 61, 5974-5978.

Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Burger, R.A., Monk, B.J. and Hampton, G.M. (2001b) Analysis of gene expression in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. U.S.A., 98, 1176-1181.

Westfall P. H.and S. S. Young. Resampling-Based Multiple Testing. John Wiley and Sons. Inc. 1993.

Weston, J. Mukherjee, S. Chapelle, O. Pontil, M. Poggio, T., Vapnik, V. Feature selection for SVMs. In *Advances in Neural Information Processing Systems 13*, Solla, S.A., Leen, T.K., Muller, K.M., eds., MIT Press, 2001.

Yeang, C.H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R.M., Angelo, M., Reich, M., Lander, E., Mesirov, J., Golub, T. (2001) Molecular classification of multiple tumor types. *Bioinformatics* 17(Suppl. 1):S316–S322.

Yeoh, Eng-Juh, Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong, & James R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell Vol 1(2) pgs 133-143, March 2002.

Yeung, K.Y., Ruzzo, W.L. (2001) An empirical study on Principal Component Analysis for clustering gene expression data. http://citeseer.nj.nec.com/yeung01empirical.html.