

## Chapter 2

# GENE EXPRESSION DATA AND SURVIVAL ANALYSIS

Peter J. Park

*Children's Hospital Informatics Program and Harvard-Partners Center for Genetics and Genomics, NRB 255, 77 Avenue Louis Pasteur, Boston, MA 02115*

**Abstract:** Finding associations between expression profiles and simple phenotypic data such as class labels has been studied extensively, including prediction algorithms for new samples based on these relationships. However, much work is needed to link expression profiles to more complex response variables, most notably survival data with censoring. Reducing the survival data to a short-term versus long-term survival indicator or using survival curves merely to demonstrate the difference between clusters of samples is not an efficient use of the data. We review some of the progress and challenges in this area. We discuss the need for more consistent results among studies done on different microarray platforms, for development of sample-specific predictive scoring schemes, and for a more comprehensive analysis that incorporates other prognostic factors and clearly demonstrates the added value of expression profiling over current protocols.

**Key words:** Cluster analysis; dimensionality reduction; censored data; Kaplan-Meier analysis; cross-platform comparisons

## 1. INTRODUCTION

From the beginning, one of the most exciting areas of application envisioned with the microarray technology has been its use in the clinic. By obtaining a ‘molecular portrait’ of diseases, we would gain fresh understanding of the disease processes at the molecular level, which would allow us to improve our classification of diseases and aid in discoveries of new subtypes. This would quickly lead, it was advertised by some, to the realization of ‘personalized medicine,’ in which diagnosis and prognosis, as

well as treatment plan, would depend on the individual's genetic information.

In the past few years, there has been a great effort in many aspects of this endeavor, to a varying degree of success. Although much is left to be desired, there has been substantial improvement in the quality of the transcript measurements, both for spotted cDNA arrays and for oligonucleotide arrays by Affymetrix. There have been some alternative technologies as well, especially the spotted or printed oligonucleotide arrays with longer probes. In terms of analysis, much of the initial work has been in associating expression data with a binary response variable such as the labels indicating cancer or normal tissue. The common tasks have been to identify genes that are highly correlated with the disease classification and then to use these genes to build a prediction scheme. Numerous methods of varying complexity have been applied to this problem. Starting with the 'signal-to-noise' metric and 'weighted-voting' prediction scheme in Golub et al. [1999], a seemingly countless number of methods from numerous disciplines has been applied to this problem, ranging from traditional statistical techniques to the latest computer-intensive techniques. Unfortunately, it is still unclear which method performs the best in general because too many methods have been applied to few relatively easy datasets, all claiming superiority against a method known to be less than optimal. A subset of these methods was subsequently expanded to the case of multiple classes, in order to deal with many subtypes of diseases [Bhattacharjee et al., 2001; Ramaswamy et al., 2001; Pomeroy et al., 2002; Rifkin et al., 2003]. Many modifications to the multi-class problem, however, have been relatively simple extensions of the binary case, in which a series of one versus many comparisons are combined.

There are other types of data besides these nominal ones that will be important in more comprehensive studies in the future. In ordinal data, the order is important in the categories, such as 'minor', 'moderate', 'severe', or 'fatal' for a disease progression; in discrete data, both the order and the magnitude are important, such as the number of relapses of a disease; and in continuous data, the measurements are not restricted to specified values. Effective methods are yet to be determined or developed in most of these cases.

An important phenotypic variable which has received more attention recently is the patient survival times. It has been gradually recognized that gene expression must be considered in the context of all other patient characteristics and that it can provide more information than simple disease classification. Survival times are obviously an important characteristic that has direct and immediate implications. Survival analysis is a collection of statistical methods for describing the distribution of survival/failure times or

any other time-to-event, and a large number of tools exist in that literature. However, there has not been a strong link yet to the high-dimensional setting encountered with expression data. Some properties and methods of analysis for survival data will be described briefly in the next section, but it suffices to note that the number of studies correlating gene expression with survival data has increased dramatically. Well represented already are different types of cancers, but now there are studies involving other clinical aspects, such as renal allograft rejection [Sarwal et al., 2003].

With survival data, careful analysis is imperative [Altman and Royston, 2000]. For example, many earlier studies may be flawed in their claims of high prediction rate in classification of samples due to a selection bias [Ambroise and McLachlan, 2002; Simon et al., 2003]. A simple simulation in Simon et al. [2003] shows that a high classification rate can be achieved in the popular leave-one-out cross-validation even for randomly generated data if the sample that is left out for testing has been used in generating the list of genes used in prediction. While this may appear obvious, the mistake happens surprisingly often. It is common, for example, to normalize the data and filter the genes to get a manageable number of genes, in the order of few hundred or a thousand genes, before the main analysis including cross-validation is carried out. The estimation of correct leave-one-out prediction rate, however, requires that the whole process including normalization and filtering be repeated each time a sample is left out. The effect of normalization while including the validation sample may not be large, but the effect of a filtering in the same way is often larger than expected. This error is no longer prevalent in the literature, but other subtle issues still remain. In some instances, the estimation of the prediction rate involves some circularity, with genes used as predictors having been used in the first place to define the groups [Sorlie et al., 2003].

## **2. CURRENT USE OF SURVIVAL DATA**

The main difficulty with patient survival data is the presence of censoring. Censoring occurs when the outcome is not observed for a patient. For example, in a cancer trial, a group of patients is followed prospectively for a period of time and the outcome variable may be time to death. At the end of the study period, however, mostly likely not all patients will have died; also, some patients may have left the study early for reasons unrelated to the disease or trial. We may know that a patient has lived at least two years, for instance, but do not know the exact time of death. Death is one example of an event; in general, the response variable can be any time-to-event data. Common variables include time to relapse of a disease, number

of occurrences of a disease, and time to disease-free state after a treatment. Censoring can be either left-censored or right-censored or both. In a prospective study, a group of patients with a particular disease may be recruited and followed. For some patients, the date of diagnosis may be known but they may be considered left-censored if the disease was contracted at some unknown time prior to the diagnosis. Unless this effect is severe or can be corrected in the analysis, however, we assume that the time of diagnosis is close to the start of disease and do not consider them as censored. Right-censoring is generally the more serious problem and cannot be ignored if unbiased estimates are to be obtained. We usually assume uninformative censoring, that the censoring is not related to the effects under investigations. For example, if a patient drops out of a study because he has moved to another location, that is considered uninformative; if he drops out due a deteriorating condition, that is not uninformative. Without this assumption, the analysis becomes more difficult if not impossible.

In most clinical studies, censoring is a serious issue that must be dealt with efficiently. Observed endpoints are desirable from the analysis point of view, but censoring inevitably occurs; it is not unusual to have more than half the patients censored in a trial. In gene expression studies, survival endpoints have not been used in an efficient manner so far. In the simplest approach, patients were roughly divided into two categories, for short-term and long-term survival. This reduced the problem into the dichotomous variable case, for which numerous methods are available. The problem with this, however, is that much information is thrown away, as may be evidenced in large within-group heterogeneity. If a two-year survival is used as a cut-off in a study, for example, a patient who survived just over two years may be put in the same group as someone who survived ten years while he is put in the different group from someone who survived just under two years.

In a more popular approach, many of the studies employed the strategy of clustering the patients according to their expression profiles first and then showing that the patients in different clusters have statistically significant differences in survival outcomes. Hierarchical clustering has been the favored clustering scheme and using Kaplan-Meier curves and log-rank tests for patient survival have been the common methods for demonstrating the differences among clusters so far. The Kaplan-Meier method is a nonparametric technique for estimating the probability that an individual survives beyond a given time; the idea behind the log-rank test is to construct a series of contingency tables for group versus survival status at each time at which a failure occurs and then to combine the information from the tables using the Mantel-Haenszel statistic.

While this approach has been fruitful in demonstrating that there is a relationship between expression profiles and survival, it is an *indirect* and inefficient use of the data. Prediction is made only in that a patient may be put in one group which, on the average, has a different survival function from the others. In a sense, survival data are used merely to verify the effectiveness of the clustering algorithm. In fact, further separation in the Kaplan-Meier curve has been used as a criterion for judging the quality of clustering algorithms at times. A more effective use of the data would be to build a predictive model that will make direct profile-specific estimates on a continuous scale. There have been some progress in this direction and some examples are mentioned in the next section.

### 3. CHALLENGES

#### 3.1 Technological limitations

One of the major problems in expression analysis has been the lack of consistency and reproducibility in the data. If, for instance, the relationship between an expression profile and its survival prediction holds only within a particular study, it is not clear how much of the conclusion was real and how much was an artifact of the analysis method. Before microarrays can be used more routinely for diagnostic or prognostic purposes, this issue of reproducibility must be better understood.

Some have suspected early on that there may be substantial difference in the results from the cDNA arrays manufactured in small-scale laboratories and from the oligonucleotide arrays, especially the high-density Affymetrix arrays with multiple 25-mer probes for each target sequence. This lack of concordance was first reported in Kuo et al. [2002] using the data from a panel of 60 cancer cell lines from the National Cancer Institute that were hybridized onto both cDNA and Affymetrix arrays. Subsequently, there have been other studies describing similar discordance for platforms including the spotted or printed oligo arrays [Yuen et al., 2002; Tan et al., 2003].

This issue is serious even within the same platform. Much of the work with clinical application has been done with Affymetrix arrays, starting with HuFL arrays and continuing with U95A-E, followed by U133A-B and U133 2.0 Plus series. However, while the basic fabrication technology has stayed the same, there have been important differences among the different generations of arrays, for example, with different number of probes in a probe set for each gene. Improvements have come most notably in the probe selection algorithms and in the calculation of summary expression measures.

In Nimgaonkar et al. [2003], it is observed that there is substantial disagreement among the Unigene matched genes between the HuFL and U95A platforms, with greater agreement when more probes are shared between the two probe sets for a gene. We have recently carried out a more extensive and quantitative comparison using a dataset in which each sample was hybridized both to U95A and to U133A (manuscript in submission). With several commonly used methods of matching genes across the arrays and preprocessing them, we were unable to reduce the dominant effect of the array type: an unsupervised clustering results in a separation by array type rather than disease type and there is substantial difference in the genes identified as differentially expressed.

Given the difficulties of comparing data even among succeeding generations of arrays within the same technology platform, it is not surprising that data generated with differences in samples, instruments, institutions, protocols, and platforms do not agree. Three prominent cases of diseases with multiple data sets are diffuse large B-cell lymphoma [Alizadeh et al., 2000; Rosenwald et al., 2002; Shipp et al., 2002]; lung carcinoma [Bhattacharjee et al., 2001; Garber et al., 2001; Beer et al., 2002; Wigle et al., 2002]; and breast cancer [Perou et al., 2000; Hedenfalk et al., 2001; Sorlie et al., 2001; West et al., 2001; van de Vijver et al., 2002; van 't Veer et al., 2002; Huang et al., 2003]. While the general conclusions of these studies are the same, specific results can vary substantially. In particular, the overlap of the marker gene lists is surprisingly small in general [Sorlie et al., 2003].

Another reason for the disagreement in the results is the different algorithms and their lack of robustness in data analysis. In Sorlie et al. [2001], genes useful for classification were determined using patient survival as the supervising variable in Significance Analysis of Microarrays [Tusher et al., 2001]; in Jenssen et al. [2002], the same dataset was analyzed using a variation on the univariate log-rank test on each gene. However, only 29 genes were common between the two lists containing 264 and 95 genes. Compared to a different data set [van 't Veer et al., 2002] that was analyzed with the occurrence of metastasis as the patient outcome, only two genes were in common between the lists with 174 and 95 genes.

This is in some respects reminiscent of the lack of reproducibility in association studies that look for common genetic variants, such as single nucleotide polymorphisms (SNPs), that contribute to disease susceptibility. In these studies, most associations claimed do not appear to be robust [Hirschhorn et al., 2002; Lohmueller et al., 2003]. In one study, a meta-analysis showed that of the 166 putative associations which have been studied three or more times, only six have been consistently replicated [Hirschhorn et al., 2002]. The underlying problem is similar in both

expression studies and genetic association studies: there are many factors that contribute the phenotype but each with only a modest contribution. Well-controlled studies with large samples and robust analysis are needed to verify the results in both cases.

There has been at least some success in comparing independent data sets. In Sorlie et al. [2003], class prediction using a variant of nearest-centroid classification method [Tibshirani et al., 2002] was performed, with one dataset as a training set and two independent datasets as testing sets. Although the number of genes shared in the informative gene lists was small and using a set of marker genes found in one study only to predict the outcomes in another does not perform well, using a set of common markers performed better and similar subtypes were observed in all cases [Sorlie et al., 2003].

## **3.2 Dealing with high-dimensionality**

Any correlative analysis of survival data with gene expression inherits all the problems associated with high-dimensional datasets in addition to the problems caused by censoring. This is a fundamentally difficult problem, and there are no simple solutions that are both mathematically rigorous and offer biologically meaningful interpretation. A large part of current analysis consists of exploratory analysis based on experience and available software.

After some initial filtering to eliminate non-expressed genes and genes with small variability, the next step is to further reduce the number of predictors to find informative genes. One approach is to use a well-known mathematical technique for dimensionality reduction. Principal component analysis and singular value decomposition are typical methods in this category [Alter et al., 2000]. While mathematically attractive, the two main disadvantages of these are that principal components or singular vectors may not highly correlated with the outcome variable and that it is difficult to assign meaning to them except in few simple cases. Sometimes the coefficients of principal components or singular vectors can be examined to determine those genes with large contributions, but usually a small subset of dominant genes does not exist. If the goal of an expression profiling project is to simply devise the most accurate prediction scheme regardless of its interpretability, such a dimensionality reduction method followed by a machine learning technique may give good results [Khan et al., 2001]. There are many machine learning methods such as neural networks and genetic algorithms, but support vector machines appear to perform especially well for that purpose [Brown et al., 2000].

For a more complex analysis involving survival phenotype, a better approach is to reduce the number of variables by grouping genes that are similar in some measure, creating what some have referred to ‘metagenes’ or ‘supergenets.’ There are many variations on this idea. Based on Tukey’s idea of compound covariates [Tukey, 1993], one can form a linear combination of genes with similar expressions with weights corresponding to the statistic from the two-sample t-test. This is the approach taken in Hedenfalk et al. [2001] and discussed further in Radmacher et al. [2002], including its relationship to the weighted-voting method [Golub et al., 1999]. In the tree-harvesting method [Hastie et al., 2001], a step-wise regression is used to select gene clusters of varying sizes that are related to the phenotype, based on the Cox proportional hazard model. The clusters may be derived, for example, from hierarchical clustering. In Rosenwald et al. [2002], Cox proportional hazard model was applied on individual genes and these were clustered into ‘signature groups.’ From this a smaller set was chosen as representative genes and averaged values of similar genes were included in a multivariate Cox model. The model was then used to compute a risk score for each patient. In this work, gene annotations were considered in grouping of the genes in addition to expression similarities and, as a result, the new reduced set of variables provides convenient biological interpretations. Multivariate Cox model was also fit in van de Vijver et al. [2002], but in this work expression profiles were reduced to an indicator variable as ‘good-prognosis’ versus ‘bad-prognosis’ signature. There are other methods of grouping genes for prediction, such as model-based clustering of genes [McLachlan et al., 2002]. In all these methods, the goal is to reduce the number of genes in a reasonable manner such that a conventional tool for survival analysis such as the multivariate Cox model can be used. A model with biological interpretation such as in Rosenwald et al. [2002] appears especially helpful.

For the purpose of prediction, approaches based on partial least squares have been explored with considerable promise. While principal component analysis has been popular in an unsupervised setting, principal components capture the variability in the gene expression space only and may not be highly correlated with the response variable. On the other hand, variable selection in linear regression chooses genes that are highly correlated with the response variable but do not account for the variability in the gene space. Partial least squares lies in between, producing a set of orthogonal linear combinations of genes that are predictive of the response while capturing the variability in the predictor space. The popularity of partial least squares has been due to its adaptability in the presence of a large number of variables, even when it exceeds the number of cases. It appears to work well when the number of predictors exceeds the number of cases moderately, but it is not



clear how scalable this result is for extreme cases. This approach has been applied in gene expression analysis [Nguyen and Rocke, 2002a; Johansson et al., 2003; Perez-Enciso and Tenenhaus, 2003] for nominal phenotype. For the censored phenotype, partial least squares was used as a dimension reduction tool [Nguyen and Rocke, 2002b]. In Park et al. [2002], the partial least squares approach was reformulated to an equivalent problem in the generalized linear regression setting for which partial least squares was already worked out in Marx [1996]. This formulation circumvents the issue of censored data, at the cost of increased dimension in the problem, and appears to perform very well. A related approach is based on kernel Cox regression models [Li and Luan, 2003] in the framework of support vector machines. This method is based on the reformulation of support vector machine as a penalization method in function estimation, with the negative partial likelihood in the Cox model as the loss function. As in partial least squares, a large number of genes may be included in the set of potential predictors.

### **3.3 Incorporating other patient data**

While there has been much progress in showing association between expression profiles and disease subtypes or even patient survival, its practical value in the clinical setting over current protocols and guidelines has not been demonstrated as convincingly in most instances. This may be one reason for the absence of microarray experiments in the clinic at this point, even after numerous studies over many years claiming the usefulness of expression profiling.

There are several reasons for this. The first is that much initial work may not have been as practical as they might have appeared at first. In some cases, it is not surprising that certain types of tumors can be distinguished with expression data, since gene expression simply reflects the features of the different cell type or other underlying characteristics. Sometimes the main distinguishing feature of expression profiles in different groups may reflect a mutation in a gene, such as in breast cancer, in which case a screening for the mutation directly would be more cost-effective and as accurate. A main conclusion of Hedenfalk et al. (2001), for example, was that mutations of BRCA1 and BRCA2 influence the expression of a group of genes. In other cases, expression profiling has not been shown conclusively to be better than immunohistochemical staining that are easier to perform and less expensive.

Even when gene expression patterns are useful for classification of disease types and stages, only a small number of such studies have

demonstrated that it is superior to current set of clinical parameters. Often, to show clinical relevance, a subclass of patients considered to be in the same stage of a disease under a standard protocol is shown to have varying survival times depending on their expression profiles. Substantial heterogeneity within a same category implies that further refinement using expression may be desirable. For example, in the case of diffuse large B-cell lymphoma patients, those with similar International Prognostic Index (IPI) still showed significantly different Kaplan-Meier curves when classified based on their expression profiles [Rosenwald et al., 2002]. In breast cancer, those patients with the same lymph-node status or risk group status showed significant differences in expression profiles with respect to both metastasis-free period and overall survival [van de Vijver et al., 2002]. This evidence of heterogeneity within a same group provides strong evidence; however, it may be, for example, that heterogeneity also exists in terms of the existing clinical parameters within those grouped by expression profiles.

To clearly demonstrate that expression profiling indeed contributes to a better classification and prediction after the effect of other prognostic factors are accounted, a multivariate model with other potential predictors should be considered. This was done, for example, in van de Vijver et al. [2002], although expression signature is entered only as an indicator variable in that work. Another method is to have a large enough cohort within a single stratum of patients with similar characteristics. In metastatic renal cell cancer, the current prognostic indicators are stage, grade, and Eastern Cooperative Oncology Group status and among stage IV tumors, no clinical parameters exist for predicting time to failure [Vasselli et al., 2003]. By considering only similarly staged patients with no other known prognostic indicator, clinical relevance of expression profiling was evidenced clearly in Vasselli et al. [2003]. More careful, integrative analysis in similar directions would be an important contribution.

## 4. CONCLUSIONS

We have briefly reviewed the use of survival data in the context of analyzing and utilizing gene expression data. While distinct expression profiles have been correlated with many disease types and this has resulted in much insight into the biological mechanism underlying these diseases, a more direct way to demonstrate their usefulness for patient care is by linking them to patient survival. Much of the work so far, however, has not utilized the survival data efficiently. In some cases, the survival data were used simply to divide the patient samples into short-term and long-term survival groups, so that previous methodologies for binary classification and

prediction can be used; in other cases, the survival information was used merely to demonstrate that the groups obtained through a clustering method were sufficiently different. In order to take full advantage of the expression data, it is important to develop new statistical methodologies that are suited for survival data analysis in the high-dimensional setting, especially in the area of effective prediction algorithms involving censored data. Simple validation techniques also need to be developed, similar to the  $n$ -fold cross-validation approach that dominates the expression literature. It is also important to carry out careful analysis to demonstrate not simply that expression profiles are correlated with survival but that they are valuable when added to the information contained in more mundane covariates and currently available prognostic factors.

We have focused mostly on dealing with the case of censored response variable here, but there are more difficult cases that will become important in future studies. In some cases, the questions have been addressed already in the survival analysis or the clinical trials literature and need to be modified for use with genomic data; in other cases, new methods need to be developed to answer fresh questions. When microarrays become part of longitudinal studies, methodologies will be needed to deal with repeated measurements [Laird and Ware, 1982]. Also, there may be more than a single phenotypic response in future studies and methods for multivariate response variables need to be studied. Incorporating other genomic data other than microarrays effectively also remains an issue. Finally, it is important in these cases that the more traditional statistical approach with emphasis on model building followed by model-checking with residuals and outlier detection needs to be reconciled with the more algorithmic approach driven by prediction rates and functional minimization from the computer science community.

## **5. ACKNOWLEDGEMENTS**

I would like to thank the three anonymous referees for their careful reading of the manuscript and helpful suggestions.

## **6. REFERENCES**

Alizadeh AA, Eisen MB, Davis RE, Ma C, Losses IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Martl GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM.

- (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503-11
- Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97:10101-6
- Altman DG, Royston P (2000) What do we mean by validating a prognostic model? *Stat Med* 19:453-73
- Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* 99:6562-6
- Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 8:816-24
- Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* 98:13790-5
- Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* 97:262-7
- 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 (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci US A* 98:13784-9
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-7
- Hastie T, Tibshirani R, Botstein D, Brown P (2001) Supervised harvesting of expression trees. *Genome Biol* 2:RESEARCH0003
- Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J (2001) Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 344:539-48
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. *Genet Med* 4:45-61
- Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT (2003) Gene expression predictors of breast cancer outcomes. *Lancet* 361:1590-6
- Jenssen TK, Kuo WP, Stokke T, Hovig E (2002) Associations between gene expressions in breast cancer and patient survival. *Hum Genet* 111:411-20
- Johansson D, Lindgren P, Berglund A (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics* 19:467-73
- Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 7:673-9
- Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 18:405-12

- Laird NM, Ware JH (1982) Random-effects models for longitudinal data. *Biometrics* 38:963-74
- Li H, Luan Y (2003) Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pac Symp Biocomput*:65-76
- Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 33:177-82
- Marx B (1996) Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics* 38:374-381
- McLachlan GJ, Bean RW, Peel D (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 18:413-22
- Nguyen DV, Rocke DM (2002a) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* 18:1216-26
- Nguyen DV, Rocke DM (2002b) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics* 18:1625-32
- Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, Kohane IS (2003) Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics* 4:27
- Park PJ, Tian L, Kohane IS (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics* 18 Suppl 1:S120-7
- Perez-Enciso M, Tenenhaus M (2003) Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum Genet* 112:581-92
- Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D (2000) Molecular portraits of human breast tumours. *Nature* 406:747-52
- Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415:436-42
- Radmacher MD, McShane LM, Simon R (2002) A paradigm for class prediction using gene expression profiles. *J Comput Biol* 9:505-11
- Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 98:15149-54
- Rifkin R, Mukherjee S, Tamayo P, Ramaswamy S, Yeang CH, Angelo M, Reich M, Poggio T, Lander ES, Golub TR, Mesirov J (2003) An Analytical Method For Multi-class Molecular Cancer Classification. *SIAM Review* 45:706-723
- Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 346:1937-47
- Sarwal M, Chua MS, Kambham N, Hsieh SC, Satterwhite T, Masek M, Salvatierra O, Jr. (2003) Molecular heterogeneity in acute renal allograft rejection identified by DNA microarray profiling. *N Engl J Med* 349:125-38
- Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002) Diffuse large B-cell lymphoma

- outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 8:68-74
- Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 95:14-8
- 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 (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 98:10869-74
- Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 100:8418-23
- Tan PK, Downey TJ, Spitznagel EL, Jr., Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 31:5676-84
- Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* 99:6567-72
- Tukey JW (1993) Tightening the clinical trial. *Control Clin Trials* 14:266-85
- Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98:5116-21
- van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347:1999-2009
- van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530-6
- Vasselli JR, Shih JH, Iyengar SR, Maranchie J, Riss J, Worrell R, Torres-Cabala C, Tabios R, Mariotti A, Stearman R, Merino M, Walther MM, Simon R, Klausner RD, Linehan WM (2003) Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proc Natl Acad Sci U S A* 100:6958-63
- West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Jr., Marks JR, Nevins JR (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* 98:11462-7
- Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, Lu C, Woodgett J, Seiden I, Johnston M, Keshavjee S, Darling G, Winton T, Breitkreutz BJ, Jorgenson P, Tyers M, Shepherd FA, Tsao MS (2002) Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res* 62:3005-8
- Yuen T, Wurbach E, Pfeffer RL, Ebersole BJ, Sealfon SC (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res* 30:e48



<http://www.springer.com/978-0-387-23074-0>

Methods of Microarray Data Analysis IV

Shoemaker, J.S.; Lin, S.M. (Eds.)

2005, XVI, 256 p., Hardcover

ISBN: 978-0-387-23074-0