Identification of Prognostic Genes, Combining Information across Different Institutions and Oligonucleotide Arrays

Jeffrey S. Morris, Guosheng Yin, Keith Baggerly, Chunlei Wu, and Li Zhang

University of Texas MD Anderson Cancer Center
1515 Holcombe Blvd, Box 447
Houston, TX 77030-4009
(713) 794-1720
jeffmo@odin.mdacc.tmc.edu

ABSTRACT

Our objective in this analysis is to identify genes whose expression levels are correlated with survival for patients with lung adenocarcinoma. Identification of such genes may lead to better clinical prognosis and identification of higher risk groups for whom more aggressive treatment may be appropriate. This goal requires us to combine data across the Harvard and Michigan studies, which used two different versions of Affymetrix oligonucleotide microarrays. We combined information across different chip types by identifying common probes on the two chips, and then combining them together into "probesets" based on UNIGENE clusters. We quantified the expression levels of each probeset using a free-energy model of binding interactions on oligonucleotide arrays that decomposes the observed probe signals in terms of the effects of gene-specific and generic non-specific binding. We used multivariable Cox regression models to identify genes providing prognostic information on patient survival *above and beyond* standard clinical predictors. We were able to identify a set of 26 genes that appear to be prognostic, many of which appear interesting and warrant future investigation.

Keywords: Cox Regression, Meta-analysis, NSCLC, Oligonucleotide Microarrays.

1. INTRODUCTION

Our analytical objective is to identify prognostic genes for lung adenocarcinoma patients, combining information from studies performed at different institutions using different microarray technologies. Our focus is on genes whose expression levels are correlated with patient survival and offer prognostic information beyond that provided by known clinical factors, e.g. stage of disease. We considered all four data sets provided, but chose to focus on the Harvard [3] and Michigan data sets [1] because the survival data for adenocarcinoma patients in the Toronto data set were not mature (only 3 events in 18 patients) and the Stanford data contained a large proportion of metastatic patients (12/30), while the other studies did not. We thus considered data from 211 patients, 125 from the Harvard study and 86 from Michigan.

2. ANALYTICAL METHODS

2.1 Combining Information across Institution

We first compared the clinical variables in the Harvard and Michigan studies to see if the patient populations were similar. We found the two studies had comparable distributions of age, gender, and smoking status (p>0.05 for all). The stage distributions were slightly different, since the Michigan study contained only stage 1 and stage 3 cancers (67 and 19, respectively), while the Harvard study contained patients at all 4 stages (76, 23, 11, and 15, respectively). However, the proportions of advanced (stage 3 and 4) vs. local (stage 1 and 2) disease were similar in the two groups (0.22 vs. 0.78 for Michigan, 0.21 vs. 0.79 for Harvard, p>0.05). The mean follow-up time in the two studies was 38 and 39 months, respectively. In spite of these similar characteristics, the patients in these two studies demonstrated significantly different survival distributions, with the Harvard patients tending to have worse prognoses. Figure 1 contains the Kaplan-Meier plots for these two groups. This difference was statistically significant (p=0.005, Cox model) even after adjusting for age and stage, so we decided to include a fixed institution effect in all subsequent survival modeling to account for apparent differences in the

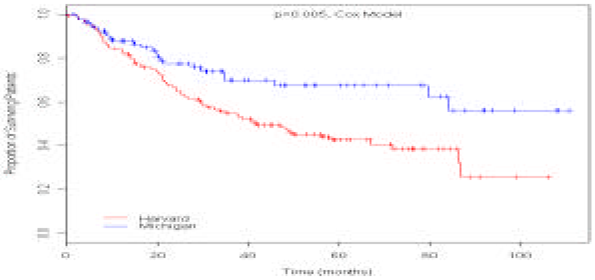

Figure 1. Kaplan-Meier Plots for Harvard and Michigan Studies. The p-value corresponds to the institution factor in a multivariable Cox model which also includes age and stage of disease

patient populations for these two studies. In spite of the difference in survival distributions, the two patient populations seemed similar enough that we determined it was reasonable to pool them together in a common analysis.

2.2 Combining Information across Different Oligonucleotide Arrays using "Partial Probesets"

A major challenge in pooling these studies is that they used different versions of the Affymetrix oligonucleotide chip for their microarray analyses. The Michigan study used the HuGeneFL Affymetrix chip, containing 6,633 probesets, each with 20 probe pairs, while the Harvard study used the newer HG_U95Av2 chip with 12,625 probesets, each with 16 probe pairs. It is not possible to quantify the expression levels simply using the Affymetrix-determined probesets when combining data across chip types, since some genes are present on one chip but not the other, and many genes common to both chip types use different sets of probes so their expression levels are not comparable.

In order to obtain comparable gene expression levels across the two chip types, we first identified probes that were common to both the HuGeneFL and HG_U95Av2 chips. Next, we mapped these probes to genes using the current annotation of HG_U95Av2 according to the links in the UNIGENE database. Probes belonging to the same UNIGENE IDs were grouped together to form new "probesets". We then discarded any probeset containing fewer than 3 probes, leaving us with 4,101 probesets common to each chip. 84% of the probesets contained 10 or fewer probes and the median probeset size was 7. The expression levels for each gene were subsequently quantified using these "partial probesets", i.e. probesets containing only the subset of probes present on both chip types.

2.3 Preprocessing and Quantifying Gene Expression Levels

We obtained log-scale quantifications of the gene expression levels using Li Zhang's Positional Dependent Nearest Neighbor (PDNN) model [8], which was introduced in last year's CAMDA competition [7]. This method fits a free-energy model of binding interactions on oligonucleotide arrays, which decomposes the observed probe signals in terms of the effects of gene-specific and generic non-specific binding. This method has been shown to be more accurate and reliable than MAS 5.0 (Affymetrix, Inc.) or dChip (Li C and Wong WH), using the Latin-square test data set provided by Affymetrix for calibrating MAS 5.0 [8].

The raw intensities for each microarray image were converted to the log scale and re-plotted to screen for bad chips. Several samples were flagged as outliers and removed from consideration. From the Michigan data set, samples L54, L88, L89, and L90 contained a large dark spot at the center of the chip, which was obvious in our log-scale plot, and L22, L30, L99, L81, L100, and L102 contained a large number of extremely bright outliers according to MAS5.0. For the Harvard data set, two outlier chips were detected using dChip (CL2001040304 and CL2001041716) and removed. For the Harvard samples with replicate arrays, we kept only the most recently run chip. This left us with matching clinical and microarray data for 200 patients, 124 from Harvard and 76 from Michigan.

We next removed the half of the genes with the lowest mean expression levels across all samples since these genes had expression levels too low to be considered reliable, leaving us with 2,505 genes. We then normalized the gene expression values on each chip by using a linear transformation to force each chip to have a common mean and standard deviation across genes.

2.4 Assessing our "Partial Probeset" Method

Our partial probeset method yields comparable estimates of the gene expression levels across different chip types, and thus allows a pooled analysis of expression levels coming from different chip types. However, in principle it is possible that by building probesets using smaller numbers of probes, we introduce extra variability into the gene expression quantifications that may render them less reliable.

To investigate this possibility, we compared the quantifications using our partial probesets with those from the full probesets determined by the HG_U95Av2 annotations. There was no extra variability evident in the log-expression levels when using the partial probesets versus the full probesets determined the by the HG_U95Av2 annotations, and there was strong evidence of agreement between the sample-to-sample variability for each gene using the two methods (r=0.943, p<0.0001, see Figure2a). We also observed strong agreement in the relative quantifications across samples, measured by computing the Spearman correlation between the log-quantifications for the partial and full probesets for each gene in samples from the Harvard data set (median r=0.95 across genes). We observed similar patterns in the Michigan data. It does not appear that we lose precision by using the partial probesets instead of the full ones. One possible explanation for this is that perhaps Affymetrix carried the most reliable probes forward when moving from the HuGeneFL to the HG_U95Av2.

Figure2b contains plots of these Spearman correlations vs. the standard deviation for each gene. Not surprisingly, the relative agreement between partial and full probeset quantifications was weaker for less variable genes. Based on this, we removed all genes with very small standard deviations across samples (sd<0.20) or smaller Spearman correlations (r<0.90), leaving us with 1,036 genes for consideration.

2.5 Identifying Prognostic Genes

Our goal was to identify prognostic genes offering predictive information on patient survival. We are not primarily interested in finding genes that are simply surrogates for known clinical prognostic factors like stage, since these factors are easily known without collecting microarray data. Rather, we are interested in finding genes that explain the variability in patient survival left after modeling the clinical predictors. Thus, we fit multivariable survival models, including clinical covariates in all survival models we used to identify prognostic genes.

We applied Cox regression models to the survival data combined across both institutions. Our best clinical model included age and disease stage (dichotomized as low, stage 1-2, and high, stage 3-4). Thus, we screened the 1036 genes to find potentially prognostic ones by fitting a series of multivariable Cox models containing age, stage, institution, and the log-expression of one of the genes as predictors. We obtained the exact p-values for each gene coefficient using a permutation approach. We also obtained asymptotic p-values using likelihood ratio tests (LRT) as well as bootstrap-based p-values to assess robustness of our results. A small p-value for a given gene indicates potential for that gene to provide prognostic information on survival beyond the clinical covariates.

If there were no prognostic genes, statistical theory suggests that a histogram of these p-values should follow a uniform distribution. An overabundance of small p-values would indicate the presence of prognostic genes. We fit a Beta-Uniform mixture model to this histogram of p-values using a method called the Beta-Uniform Mixture method (BUM [4]), which partitions the histogram into two components – a Beta component containing the prognostic genes and Uniform component containing the non-significant ones. Various criteria can be used with this method to determine a cutpoint between these components. We chose the false discovery rate (FDR, introduced by Benjamini and Hochberg [2]), which estimates the proportion of genes flagged as prognostic that are in fact not prognostic. Given a choice for FDR, the BUM method yields a p-value cutoff below which a gene is flagged as significant.

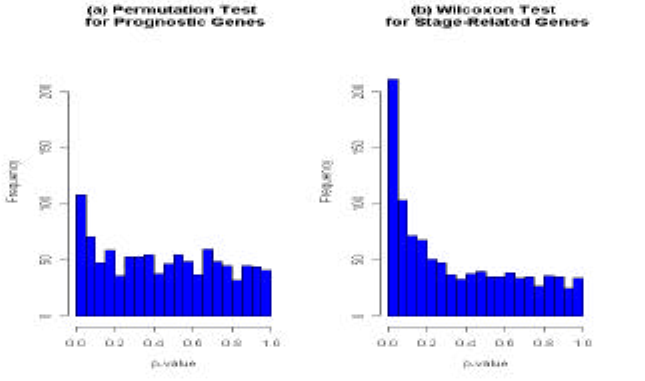

Figure 3. (a) Histogram of p-values from permutation test on gene coefficient in Cox model containing clinical covariates and each one of the 1036 candidate genes. The corresponding histogram for the LRT is nearly identical **(b)** Histogram of p-values

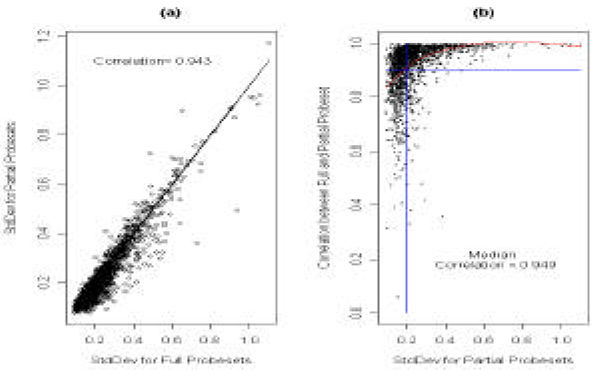

Figure 2. (a) Standard Deviation of Log-Expression Levels for Genes Quantified using full HG_U95Av2 Probesets and our "Partial Probesets" Containing only Probes Present on Both Chip Types, and (b) Spearman Correlation for Genes Quantified using Full and Partial Probesets vs. Standard Deviation of Log-Expression Levels using Partial Probesets. The red curve indicates the loess fit demonstrating the relationship between gene variability and reliability. The horizontal and vertical blue lines indicate cut points, for which genes below or to the left were removed from

2.6 Identifying Stage-Related Genes

We also identified genes differentially expressed by cancer stage by applying the BUM model to p-values from nonparametric Wilcoxon tests comparing median expression levels for early (stage 1-2) and late (stage 3-4) stage lung adenocarcinoma.

3. RESULTS

3.1 Identifying Prognostic Genes

Figure 3a contains the histogram of permutation test p-values assessing the prognostic significance of each gene. The increased number of very small p-values indicates the presence of some genes that appear linked with patient prognosis. Table 1 (next page) contains a set of 26 genes that were flagged by the BUM method using FDR<0.20, which are those genes with permutation p-values less than 0.0025. Analogous BUM analyses found 16 of these genes were also flagged based on the LRT, and 18 using the bootstrap.

3.2 Identifying Stage-Related Genes

We also identified a set of genes differentially expressed by clinical stage (early vs. late). Figure 3b contains the histogram of stage p-values, and indicates a very large number of significant genes. Using the BUM method with FDR<0.20, more than 1/3 of the genes (346/1036) were flagged as differentially expressed by stage. This is in contrast to the very small number (26) of genes flagged as prognostic. There are 71 genes flagged using FDR<0.05, which corresponds to a p-value cutoff of 0.0064. Table 2 contains the top 10 genes from this list, along with their identities. Only 1 of the 26 genes we flagged as prognostic were in the set of 71 genes flagged as related to stage using FDR<0.05 (STK25).

4. INTERPRETATION OF RESULTS

We were able to link 14 of our 26 prognostic genes to lung cancer, cancer in general, or other lung disease based on existing literature. These genes are in boldface in the table.

The top gene in our list, FCGRT, is induced by Interferon γ in treatment of SCLC. The negative sign on our coefficient indicates it is a positive prognostic factor, i.e. patients with high levels of this gene tended to have better prognosis. According to our model, every doubling of expression level of this gene corresponds to a 8-fold reduction in risk for death (hazard).

RRM1 has shown to be overexpressed in NSCLC, and one study found that NSCLC patients treated with gemcitabine/cisplatin with low RRM1 mRNA levels showed significantly longer survival times. The positive sign on the regression coefficient indicates our analysis also considers this gene a negative prognostic factor, meaning higher expression levels correspond to poorer prognosis. Every doubling of expression level corresponds with a 6-fold increase in hazard.

Overexpression of selenoprotein W, 1 (SEPW1) has been shown to markedly reduce the sensitivity to H_2O_2 cytotoxicity in NSCLC cell lines. This gene appears as a positive prognostic factor in our analysis.

FSCN1 has been demonstrated to be a prognostic marker of invasiveness in Stage 1 NSCLC, and appears as a negative prognostic factor in our analysis.

Some genes in our list are lung cancer markers, either for NSCLC (CHKL, ENO2) or SCLC (CLU, CPE). ADBRK was found to be co-expressed with Cox-2 in lung adenocarcinoma.

Some genes have been linked to other cancers. While it is possible that the connections of these genes with lung cancer are circumstantial, we mention them here because some may be interesting and turn out to be relevant to lung cancer. BCL9 is over-expressed in some cases of ALL, and NFRKB is amplified in AML. BTG2 has been demonstrated to inhibit cell proliferation in primary mouse embryo fibroblasts lacking functional p53, and is a positive prognostic gene in our analysis. ATIC is a fusion partner of ALK that defines a subtype of anaplastic large cell lymphoma (ALCL), and ALK itself has been linked with lung cancer. TPS1 is a unique protease, released from mast cell secretory granules into the respiratory tract of patients with inflammatory disease of the airways.

Only one of our 26 prognostic genes was deemed significantly associated with stage (STK25), and many flagged genes did not seem to have any association with stage. None of our genes appeared in the list of top 100 genes from the Michigan analysis [2], and we only found one (CPE) mentioned in the Harvard paper [3]. CPE was one of the genes defining a neuroendocrine cluster they identified and associated with poor prognosis.

Table 2. Top Genes Linked with Stage (using Wilcoxon Test).

UNIGENE ID	Stage p-value	Gene Identity
Hs.77665	0.000006*	KIAA0102; KIAA0102 gene product
Hs.76941	0.000016*	ATP1B3; ATPase, beta 3 polypeptide
Hs.109606	0.000022*	CORO1A; Coronin, acta binding protein, 1A
Hs.432605	0.000044*	UGCG; UDP-glucose ceramide glucosyltransferase
Hs.346918	0.000068*	PSMA3; Proteasome (prosome, macropain) subunit, alpha type 3
Hs.114360	0.000076*	TSC22; Transforming growth factor beta-stimulated protein TSC-22
Hs.2450	0.000105*	LARS2; Leucyl-tRNA synthetase, mitochondrial
Hs.168157	0.000163*	NFYC; Nuclear transcription factor Y, gamma
Hs.459	0.000193*	SLC18A3; Solute carrier family 18

Table 1. Set of genes flagged as prognostic by BUM on the permutation p-values using FDR<0.20. Also included are the LRT and bootstrap p-values, estimates of the Cox model coefficient and corresponding 99% bootstrap confidence intervals, and the p-value linking gene with stage of disease. A '*' indicates the p-value was below the BUM significance threshold. The identity of the genes is also given, with boldface type indicating we were able to find existing literature linking that gene with lung cancer, cancer in general, or other lung disease.

UNIGENE ID	Coef	99% Bootstrap CI[1]	Prognostic p-values[2]			Stage p-value[3]	Gene Identity
			Permutation	LRT	Bootstrap		
Hs.111903	-2.07	(-3.82, -0.60)	<0.00001*	0.00014*	0.0006*	0.15432	**FCGRT; Fc fragment of IgG receptor**
Hs.146580	1.46	(0.74, 2.63)	0.00001*	0.00002*	<0.0001*	0.28170	**ENO2; Enolase 2**
Hs.374357	-2.81	(-5.55, -0.37)	0.00001*	0.00435	0.0040*	0.05819	**NFRKB; Nuclear factor related to kappa B binding**
Hs.2934	1.81	(0.53, 3.13)	0.00002*	0.00008*	<0.0001*	0.32059	**RRM1; Ribonucleotide reductase M1 polypeptide**
Hs.343564	-2.35	(-4.33, -0.80)	0.00004*	0.00069*	0.0006*	0.01249	TBCE; Tubulin-specific chaperone e
Hs.181013	1.92	(0.39, 3.65)	0.00008*	0.00020*	0.0004*	0.57611	Similar to phosphoglycerate mutase 1
Hs.90280	1.81	(0.07, 3.73)	0.00009*	0.00153*	0.0004*	0.77072	**ATIC; IMP cyclohydrolase**
Hs.154886	-1.43	(-3.71, 0.18)	0.00010*	0.02305	0.0260	0.97865	**CHKL; Choline kinase-like**
Hs.380774	-2.37	(-4.49, -0.70)	0.00017*	0.00012*	0.0002*	0.80504	DDX3; DEAD/H box polypeptide 3
Hs.34789	-1.64	(-2.78, -0.57)	0.00020*	0.00010*	0.0010*	0.10831	OST; oligosaccharyltransferase
Hs.75360	0.72	(0.20, 1.30)	0.00031*	0.00053*	0.0010*	0.08839	**CPE; Carboxypeptidase E**
Hs.83636	-2.20	(-4.26, -0.60)	0.00044*	0.00678	0.0030*	0.48465	**ADRBK1; Adrenergic, beta, receptor kinase 1**
Hs.122607	-1.64	(-3.48, 0.22)	0.00067*	0.03602	0.0460	0.05663	**BCL9; B-cell CLL/lymphoma 9**
Hs.155291	1.33	(0.05, 1.56)	0.00068*	0.00279*	0.0006*	0.05472	BZW1; Basic leucine zipper and W2 domains 1
Hs.334455	-0.64	(-1.32, -0.07)	0.00106*	0.00217*	<0.0001*	0.88180	**TPS1; Tryptase, alpha**
Hs.75106	-0.52	(-0.98, -0.05)	0.00109*	0.00239*	0.0024*	0.01439	**CLU; Clusterin**
Hs.168669	-2.19	(-4.22, -0.22)	0.00118*	0.00405	0.0020*	0.65340	OGDH; Oxoglutarate dehydrogenase
Hs.155206	2.29	(0.22, 4.72)	0.00122*	0.00152*	0.0080	0.00482*	STK25; Serine/threonine kinase 25
Hs.21413	-1.70	(-3.83, 0.34)	0.00143*	0.00988	0.0220	0.01415	KCC2; potassium-chloride transporter 2
Hs.14231	-1.29	(-3.14, 0.02)	0.00145*	0.01026	0.0160	0.02819	**SEPW1;Selenoprotein W, 1**
Hs.118400	0.66	(-0.02, 2.93)	0.00150*	0.00241*	0.0103	0.08244	**FSCN1; Fascin homolog 1, actin-bundling protein**

5. DISCUSSION

There are two key challenges when pooling information across multiple microarray studies performed at different institutions. First, one must somehow account for institution and study-level differences using principles of meta-analysis. Stangl [5] describes a Bayesian hierarchical approach to this problem based on exponentially distributed survival times, and Therneau and Grambsch [6] (chapter 9) describe frailty models, a frequentist approach to this problem that can be applied to Cox models. These methods are not practical to implement given just two institutions, but our approach of a fixed offset for institution has roughly the same effect, adjusting for the institution heterogeneity in baseline risk. Given more sites, we would recommend using frailty models or Bayesian hierarchical models to account for this heterogeneity.

Second, one must find a way to combine data across microarray platforms. We devised a new method applicable to oligonucleotide arrays in which we identify probes present on both platforms, and combine them together into probesets using UNIGENE. Our experience with this data suggests that this method is effective, leading to comparable quantifications across chip type without evidence of any precision loss that one might expect. This suggests to us that this approach may be a good general method to consider, and should be further investigated.

Our biological goal in this analysis was to identify prognostic genes, meaning genes offering information on patient survival beyond that provided by known clinical predictors. We

[1] CI and p-value obtained from fitted coefficients from Cox model on 10,000 bootstrap datasets, obtained by randomly selecting 200 samples with replacement and retaining the sample-gene link.

[2] The prognostic p-values from various methods. The permutation p-values were obtained from fitted Cox coefficients on 100,000 datasets with gene expression values randomly permuted across samples. The LRT p-values were obtained from Likelihood Ratio Test, assuming asymptotic Chi-squared distribution with 1 degree of freedom. The bootstrap p-values were obtained using the 10,000 bootstrap samples.

[3] Obtained using Wilcoxon test comparing median expression levels for patients with early (stage 1-2) and late (stage 3-4) stage cancer.

accomplished this by fitting multivariable Cox models containing the clinical predictors in addition to the genes. We feel that this is crucial, since a gene whose expression is simply a surrogate for known clinical predictors does not seem nearly so useful to us since we can build a prognostic model using clinical covariates without the additional time and expense required to collect microarray data. While this type of multivariable analysis may result in fewer prognostic genes, we feel that this list has the potential to be more interesting biologically because we know that the genes we flag have the potential to explain the variability in patient survival not already explained by the clinical predictors. Many of the genes in our short list seem biologically interesting, as there is some evidence from existing literature that they are related to lung cancer.

There are various assumptions we made in our modeling – the proportional hazards assumption inherent in the Cox modeling assumes that the increase in risk due to the genes is stationary over time. While it was not possible to rigorously check this assumption for all models, we did spot checks for some of our models using diagnostics based on weighted residuals [6], and found that this assumption seemed reasonable for the cases we checked. In addition, our model assumes that the effect of the gene is linear in the log-scale, i.e. the risk is increased by a constant factor for every doubling of expression level. This approach should be robust enough to detect many strong associations between expression level and survival, even in cases for which the relationship is not strictly linear in the log scale.

There are many methods in the existing literature for determining which genes are differentially expressed while accounting for the multiplicity problem inherent to the microarray platform. We used the BUM method, which is very flexible and easy to use. It has the advantage that being based on p-values, it can be used with any valid statistical test. However, this method also depends heavily on the underlying assumptions of the hypothesis test used, so it is important to either verify the assumptions or use robust methods like nonparametric or permutation-based tests to get the p-values. Also, the BUM assumes independence across genes, which is used as a working model in many other methods, as well, but is not entirely true in practice. Studies need to be done to assess the effect of this independence assumption when it does not hold.

6. CONCLUSION

We have introduced a method based on partial probesets that appears to be effective for combining expression data from different oligonucleotide arrays. Using this method, we have combined information across the Harvard and Michigan studies and identified a set of genes that appear to be prognostic for lung adenocarcinoma, providing information above and beyond known clinical predictors. Some of these genes appear to be biologically interesting and are worthy of future consideration.

7. ACKNOWLEDGEMENTS

We would like to thank David Stivers and Kevin Coombes for helpful discussions regarding this analysis. We would also like to thank Lianchun Xiao and Sang-Joon Lee for their contributions on this project.

8. REFERENCES

[1] Beer, D, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 9 (816), 2002.

[2] Benjamini, Y, and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing”. *Journal of the Royal Statistical Society, Series B*, 57(1), 289-300, 1995.

[3] Bhattacharjee, A, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS* 98 (24), 13790-13795, 2001.

[4] Pounds, S and Morris, S. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values”. *Bioinformatics*, 19, 1236—1242, 2003.

[5] Stangl, DK. Hierarchical Analysis of Continuous-Time Survival Models. *Bayesian Biostatistics*, DA Berry and DK Stangl, eds., Marcel Dekker, New York: 429-450, 1996.

[6] Therneau, TM and Grambsch, PM. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

[7] Zhang, L, Coombes, K, and Xia, L. Quantifications of Cross Hybridization on Oligonucleotide Microarrays. *Methods of Microarray Data Analysis III,* 2003.

[8] Zhang, L, Miles, MF, Aldape, KD. A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology* 21(7): 818-21, 2003.