Chapter 1

POOLING INFORMATION ACROSS DIFFERENT STUDIES AND OLIGONUCLEOTIDE CHIP TYPES TO IDENTIFY PROGNOSTIC GENES FOR LUNG CANCER

Jeffrey S. Morris, Guosheng Yin, Keith Baggerly, Chunlei Wu, and Li Zhang
The University of Texas, MD Anderson Cancer Center, 1515 Holcombe Blvd, Box 447, Houston, TX, 77030-4009

Abstract: Our goal in this work is to pool information across microarray studies conducted at different institutions using two different versions of Affymetrix chips to identify genes whose expression levels offer information on lung cancer patients' survival above and beyond the information provided by readily available clinical covariates. We combine information across chip types by identifying “matching probes” present on both chips, and then assembling them into new probesets based on Unigene clusters. This method yields comparable expression level quantifications across chips without sacrificing much precision or significantly altering the relative ordering of the samples. We fit a series of multivariable Cox models containing clinical covariates and genes and identify 26 genes that provide information on survival after adjusting for the clinical covariates, while controlling the false discovery rate at 0.20 using the Beta-Uniform mixture method. Many of these genes appear to be biologically interesting and worthy of future investigation. Only one gene in our list has been mentioned in previously published analyses of these data. It appears that the increased statistical power provided by the pooling is key to finding these new genes, since only 9 out of the 26 genes are detected when we apply these methods to the two data sets separately, i.e., without pooling.

Key words: Cox regression; Meta-analysis; NSCLC; Oligonucleotide microarrays.
1. INTRODUCTION

The challenge of this CAMDA competition was to pool information across studies to yield new biological insights, improving medical care and leading to a better understanding of lung cancer biology. We selected adenocarcinoma, since most of the available data is from this type of histology, and it is most prevalent in the general population, and we decided to focus on the survival outcome. We chose to focus our efforts on the Michigan and Harvard studies. Both studies used Affymetrix oligonucleotide arrays, but they used different versions of Affymetrix chips: the Michigan study used the HuGeneFL while Harvard used the U95Av2.

Our first goal in this work is to pool the data across different studies to identify prognostic genes for lung adenocarcinoma. By prognostic genes, we mean those whose expression levels offer information on patient survival over and above the information already provided by known clinical predictors. We predict that by actually pooling the data as opposed to merely pooling the results, we will have more statistical power to detect prognostic genes. Accomplishing this goal requires us to develop methodology to pool information across different versions of Affymetrix chips in such a way that we obtain comparable expression levels across the different chip types.

2. ANALYTICAL METHODS

2.1 Pooling Information across Studies

Before pooling the studies, we check to see if they have comparable patient populations, and we find comparable distributions of age, gender, smoking status, and follow-up time in the studies (p>0.05 for all). The stage distributions are slightly different, since the Michigan study contains only stage I and stage III cancers (67 and 19, respectively), while the Harvard study contains patients at all 4 disease stages (76, 23, 11, and 15, respectively). However, the proportions of advanced (stage III and IV) versus local (stage I and II) disease are similar in the two groups (0.22 vs. 0.78 for Michigan, 0.21 vs. 0.79 for Harvard, p>0.05). In spite of these similar characteristics, the patients in these two studies demonstrate 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 is statistically significant (p=0.005, Cox model) even after adjusting for age and stage, so include a fixed institution effect in all subsequent survival modeling to account for apparent differences in the patient populations for these two studies. In spite of the difference in survival distributions, the two patient populations seem similar enough that it is reasonable to pool the data for a common analysis.

2.2 Pooling Information across Different Oligonucleotide Arrays using “Partial Probesets”

A major challenge in pooling these studies is that different versions of the Affymetrix oligonucleotide chip were used in the microarray analyses. The Michigan study used the HuGeneFL Affymetrix chip. This chip contains 6,633 probesets, each with 20 probe pairs. By contrast, the Harvard study used the newer U95Av2 chip. This chip contains 12,625 probesets, each with 16 probe pairs. This difference in chip types raises two problems. First, some genes may be represented on one chip but not the other. Second, genes present on both chips may be represented by different sets of probes on the two chips. Since the two chip types do not contain the same probesets, we don’t expect standard analyses on these Affymetrix-determined probesets to yield comparable expression level quantifications across chips. However, there are some probes that both chips share in common, which we call “matching probes”. These probes share common chemical properties on the two chips, and so should yield comparable intensities across the two chip types. Our method focuses on these matching probes.

Our first step is to identify the matching probes present on both the HuGeneFL and U95Av2 chips. We next recombine these probes into new
probesets using the current annotation of U95Av2 based on Unigene build 160. We refer to these recombined probesets as “partial probesets”. Note that because they are explicitly based on Unigene clusters, these probesets will not precisely correspond to the Affymetrix-determined probesets. Frequently, multiple Affymetrix probesets map to the same Unigene cluster. We then eliminate any probesets consisting of just one or two probes, because we expect the summaries from these probesets to be less precise. This left us with 4,101 partial probesets. Most of the probesets (84%) of the probesets contained 10 or fewer probes and the median probeset size was 7. We had several probesets that contained more than 20 probes.

2.3 Preprocessing and Quantifying Gene Expression Levels

We convert the raw intensities for each microarray image to the log scale and re-plot them to check if there are any poor-quality arrays. We remove from consideration several arrays that have apparent quality problems. From the Michigan data set, samples L54, L88, L89, and L90 contain a large dead spot at the center of the chip, which is obvious when looking at our log-scale plot, shown in Figure #2. These dead spots may be bubbles caused by inadequate hybridization from using less than the specified 200ml of hybridization fluid. Samples L22, L30, L99, L81, L100, and L102 contain a large number of extremely bright outliers according to MAS5.0. For the Harvard data set, two outlier chips are detected using dChip (CL2001040304 and CL2001041716) and removed. For the Harvard samples with replicate arrays, we keep only the most recently run chip. The remaining data is matching clinical and microarray data for 200 patients, 124 from the Harvard study and 76 from the Michigan study.

Figure 1-2. Log intensity plot for four Michigan samples (L54, L88, L89, and L90, respectively) with inadequate hybridization in the middle of the chips.
For each patient, we obtain log-scale quantifications of the gene expression levels for each partial probeset using the Positional Dependent Nearest Neighbor (PDNN) model. This method was introduced in last year’s CAMDA competition (Zhang, Coombes, and Xia, 2003), and uses probe sequence information to predict patterns of specific and nonspecific hybridization intensities. By explicitly using the sequencing information, this model is able to borrow strength across probe sets while doing the quantification. This method has been shown to be more accurate and reliable than MAS 5.0 (Affymetrix, Inc.) or dChip (Schadt, Li, Ellis, and Wong, 2001), using the Latin-square test data set provided by Affymetrix for calibrating MAS 5.0 (Zhang, Miles, and Aldape, 2003).

We also perform other preprocessing steps. We remove the half of the probesets with the lowest mean expression levels across all samples, then normalize the log expression values by using a linear transformation to force each chip to have a common mean and standard deviation across genes. We next remove the probesets with the smallest variability across chips (standard deviation <0.20), since we consider them unlikely to be discriminatory and more likely to be spuriously flagged as prognostic. Finally, we remove the probesets with poor relative agreement (<0.90) between the partial probeset and full probeset quantifications (see Section 3). After this preprocessing, 1036 probesets remain and are considered in our subsequent analyses.

2.4 Identifying Prognostic Genes

Our main goal is 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 available without collecting microarray data. Rather, we are interested in finding genes that explain the variability in patient survival that remains after modeling the clinical predictors. Thus, we fit multivariable survival models, including clinical covariates in all survival models we use to identify prognostic genes.

We apply Cox regression models to the survival data combined across both institutions. Our best clinical model includes age and disease stage (dichotomized as low, stages I-II, and high, stages III-IV). Smoking status is only marginally significant for survival; therefore, we remove it from the model. Thus, we screen 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 obtain the exact p-values for each gene’s coefficient using a permutation approach. In this approach, we first generated 100,000 datasets by randomly
permuting the gene expression values across samples while keeping the clinical covariates fixed. Subsequently, we obtain the permutation p-value for each gene by counting the proportion of fitted Cox coefficients that are more extreme than the coefficient for the true dataset. We also obtain p-values using asymptotic likelihood ratio tests (LRT) and the bootstrap to assess robustness of our results. The results were generally concordant; see Section 4. 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 are no prognostic genes, statistical theory suggests that a histogram of these p-values should follow a uniform distribution. An overabundance of small p-values indicates 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, Pounds and Morris, 2003), 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 along with this method to determine a cutpoint between these components. We use the false discovery rate (FDR, Benjamini and Hochberg, 1995), 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.

We also identify genes differentially expressed by cancer stage by applying the BUM model to p-values from nonparametric Wilcoxon tests comparing median expression levels for early- (stages I-II) and late-stage (stages III-IV) lung adenocarcinoma.

3. ASSESSING “PARTIAL PROBESET” METHOD

Before analyzing the microarray data to identify prognostic genes, we assess whether our method for combining information across different Affymetrix chip types performs acceptably. First, we check whether the expression levels are indeed comparable across chip types. Figure 3 contains plots of the median and median absolute deviation (MAD) log expression level for each partial probeset across the Michigan samples run on the HuGeneFL chip against those from the Harvard samples run on the U95Av2 chip. The concordance between these values is 0.961 for the median and 0.820 for the MAD, so it appears that our method yields reasonably comparable expression levels across the two chips.
Figure 1-3. Median (a) and Median Absolute Deviation (b) expression levels for each partial probeset based on the Harvard samples run on the U95Av2 chips vs. the Michigan samples run on the HuGeneFL chip. The high concordance in these measures suggests we obtain reasonably comparable expression levels by using the matched probes.

Recall that our method uses only the matching probes, while completely ignoring expression level information for the non-matching probes. This means that our probesets are generally smaller than the Affymetrix-defined probesets. The median size of our “partial probesets” is 7, while the Affymetrix-defined probesets for the HuGeneFL and U95Av2 chips have 20 and 16 probes, respectively. Since additional probes can increase the precision in measuring the expression level of the corresponding gene, one might expect a loss of precision when using the partial probesets to quantify expression levels. To investigate this possibility, we quantify the expression levels for the full probesets of the Harvard samples using the PDNN model. The full probesets consist of all probes on the array mapping to the Unigene cluster, i.e., not just the matching ones. We plot the standard deviation for each gene using the full probeset versus the standard deviation for the partial probeset, given in Figure #4. If the partial probeset quantifications were considerably less precise, we would expect measurement error to cause the standard deviation to be larger for the partial probesets. There is no evidence of significant precision loss in this plot, as there is strong agreement between the standard deviations for each gene using the two methods (concordance=0.942). This may seem surprising at first, but upon further thought is reasonable, since we expect that the probes Affymetrix chooses to retain in formulating new chips may be in some sense the “best” ones.
1. POOLING INFORMATION ACROSS DIFFERENT STUDIES AND OLIGONUCLEOTIDE CHIP TYPES TO IDENTIFY PROGNOSTIC GENES FOR LUNG CANCER

Figure 1-4. Standard deviation across Harvard samples for each gene based on full and partial probesets. A “full probeset” contains all probes on the U95Av2 chip mapping to a unique Unigene ID, while the corresponding “partial probeset” contains only the subset of probes contained on both the U95Av2 and HuGeneFL chips.

We compute Spearman correlations between the partial and full probeset quantifications for each probeset to confirm that our method preserved the relative ordering of the samples, i.e., the ranks. For example, we expect that a sample with the largest expression level for a given gene using the full set of probes will also demonstrate the largest expression level for that gene when using only the matched probes. The median Spearman correlation across all probesets is 0.95, suggesting that our method does a good job of preserving the relative ordering of the samples. Interestingly, but not surprisingly, most of the lower Spearman correlations occur for probesets with less heterogeneous expression levels across samples and/or probesets containing smaller numbers of probes. Thus, it appears that our partial probeset method works quite well. We expect it to perform even better if it is used to combine information across U95 and U133 chips, since these chips share more probes in common than the HuGeneFL and U95 chips.

4. RESULTS

Figure #5(a) contains the histogram of permutation test p-values assessing the prognostic significance of each gene. The overabundance of probesets with very small p-values indicates the presence of some genes providing information on patient prognosis beyond what is offered by the modeled clinical factors. Table 1 contains a set of 26 genes that are flagged by the
BUM method using FDR<0.20, which are those genes with permutation p-values less than 0.0025. Our analogous BUM analyses find that 16 of these genes are also flagged based on the LRT, and 18 using the bootstrap. We also identify a set of genes that appear to be differentially expressed by clinical stage (early vs. late). Figure #6(b) contains the histogram of stage p-values from the Wilcoxon test, with the extreme right skewness indicating 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 with the same settings. This is not surprising, since one might expect that it is easier to identify genes related to an easily identifiable biological factor like stage than to predict how long the patient will live. There are 71 genes flagged using FDR<0.05, which corresponds to a p-value cutoff of 0.0064. Only 1 of the 26 genes we flag as prognostic is in the set of 71 genes flagged as related to stage using FDR<0.05 (STK25).

![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.](image)

![Histogram of p-values from Wilcoxon test comparing median expression levels for early and late stage cancers.](image)

**Figure 1-5.** (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 from Wilcoxon test comparing median expression levels for early and late stage cancers.

5. **INTERPRETATION OF RESULTS**

We are able to link 10 of our 26 prognostic genes to lung cancer based on the existing literature. Four others can be linked to cancer in general or other lung disease in the literature. These genes are in boldface in Table 1.
Table 1-1. Set of genes flagged as prognostic by applying BUM on the permutation p-values with FDR<0.20. Also included are the LRT and bootstrap p-values and estimates of the Cox model coefficient. 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. A negative coefficient indicates that larger expression levels of that gene correspond to a better survival outcome.

<table>
<thead>
<tr>
<th>Gene Identity</th>
<th>Coef</th>
<th>Permut</th>
<th>LRT</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCGRT; Fc fragment of IgG receptor</td>
<td>-2.07</td>
<td>&lt;0.00001*</td>
<td>0.00014*</td>
<td>0.0006*</td>
</tr>
<tr>
<td>ENO2; Enolase 2</td>
<td>1.46</td>
<td>0.00001*</td>
<td>0.00002*</td>
<td>&lt;0.0001*</td>
</tr>
<tr>
<td>NFRKB; Nuclear factor for kappaB binding</td>
<td>-2.81</td>
<td>0.00001**</td>
<td>0.00435</td>
<td>0.0040*</td>
</tr>
<tr>
<td>RRM1; Ribonucleotide reductase M1 polypeptide</td>
<td>1.81</td>
<td>0.00002*</td>
<td>0.00008*</td>
<td>&lt;0.0001*</td>
</tr>
<tr>
<td>TBCE; Tubulin-specific chaperone e</td>
<td>-2.35</td>
<td>0.00004*</td>
<td>0.00069*</td>
<td>0.0006*</td>
</tr>
<tr>
<td>Similar to phosphoglycerate mutase 1</td>
<td>1.92</td>
<td>0.00008*</td>
<td>0.00020*</td>
<td>0.0004*</td>
</tr>
<tr>
<td>ATIC; IMP cyclohydrolase</td>
<td>1.81</td>
<td>0.00009*</td>
<td>0.00153*</td>
<td>0.0004*</td>
</tr>
<tr>
<td>CHKL; Choline kinase-like</td>
<td>-1.43</td>
<td>0.00010*</td>
<td>0.02305</td>
<td>0.0260*</td>
</tr>
<tr>
<td>DDX3; DEAD/H box polypeptide 3</td>
<td>-2.37</td>
<td>0.00017*</td>
<td>0.00012*</td>
<td>0.0002*</td>
</tr>
<tr>
<td>OST; oligosaccharyltransferase</td>
<td>-1.64</td>
<td>0.00020*</td>
<td>0.00010*</td>
<td>0.0010*</td>
</tr>
<tr>
<td>CPE; Carboxypeptidase E</td>
<td>0.72</td>
<td>0.00031*</td>
<td>0.00053*</td>
<td>0.0010*</td>
</tr>
<tr>
<td>ADRBK1; Adrenergic, beta, receptor kinase 1</td>
<td>-2.20</td>
<td>0.00044*</td>
<td>0.00678</td>
<td>0.0030*</td>
</tr>
<tr>
<td>BCL9; B-cell CLL/lymphoma 9</td>
<td>-1.64</td>
<td>0.00067*</td>
<td>0.03602</td>
<td>0.0460*</td>
</tr>
<tr>
<td>BZW1; Basic leucine zipper and W2 domains 1</td>
<td>1.33</td>
<td>0.00068*</td>
<td>0.00279*</td>
<td>0.0006*</td>
</tr>
<tr>
<td>TPS1; Tryptase, alpha</td>
<td>-0.64</td>
<td>0.00106*</td>
<td>0.00217*</td>
<td>&lt;0.0001*</td>
</tr>
<tr>
<td>CLU; Clusterin</td>
<td>-0.52</td>
<td>0.00109*</td>
<td>0.00239*</td>
<td>0.0024*</td>
</tr>
<tr>
<td>OGDH; Oxoglutarate dehydrogenase</td>
<td>-2.19</td>
<td>0.00118*</td>
<td>0.00405</td>
<td>0.0020*</td>
</tr>
<tr>
<td>STK25; Serine/threonine kinase 25</td>
<td>2.29</td>
<td>0.00122*</td>
<td>0.00152*</td>
<td>0.0080*</td>
</tr>
<tr>
<td>KCC2; potassium-chloride transporter 2</td>
<td>-1.70</td>
<td>0.00143*</td>
<td>0.00988</td>
<td>0.0220*</td>
</tr>
<tr>
<td>SEPW1; Selenoprotein W, 1</td>
<td>-1.29</td>
<td>0.00145*</td>
<td>0.01026</td>
<td>0.0160*</td>
</tr>
<tr>
<td>FSCN1; Fascin homolog 1, actin-bundling protein</td>
<td>0.66</td>
<td>0.00150*</td>
<td>0.00241*</td>
<td>0.0103*</td>
</tr>
<tr>
<td>MRPL19; Mitochondrial ribosomal prot L19</td>
<td>1.12</td>
<td>0.00211*</td>
<td>0.03213</td>
<td>0.0340*</td>
</tr>
<tr>
<td>ALDH9; Aldehyde dehydrogenase 9 family</td>
<td>-1.18</td>
<td>0.00223*</td>
<td>0.00378*</td>
<td>0.0020*</td>
</tr>
<tr>
<td>PFN2; Profilin 2</td>
<td>0.63</td>
<td>0.00248*</td>
<td>0.00351*</td>
<td>0.0020*</td>
</tr>
<tr>
<td>BTG2; BTG family, member 2</td>
<td>-0.75</td>
<td>0.00232*</td>
<td>0.00580</td>
<td>0.0140*</td>
</tr>
</tbody>
</table>
The top gene in our list, FCGRT, is induced by Interferon $\gamma$ in the treatment of SCLC (Pujol, et al, 1993). The negative sign on our coefficient indicates that it is a positive prognostic factor, i.e., patients with high levels of this gene tend to have better prognoses. According to our model, every doubling of expression level of this gene corresponds to an 8-fold reduction in risk for death (hazard). RRM1 has been shown to be overexpressed in NSCLC, and one study found that patients with NSCLC who are treated with gemcitabine/cisplatin with low RRM1 mRNA levels show significantly longer survival times (Rosell, et al, 2003). The positive sign on the regression coefficient indicates that our analysis also considers this gene to be a negative prognostic factor, meaning that higher expression levels correspond to a poorer prognosis. Every doubling of the expression level corresponds with a 6-fold increase in the hazard.

Overexpression of selenoprotein W, 1 (SEPW1) has been shown to markedly reduce the sensitivity to $H_2O_2$ cytotoxicity in NSCLC cell lines (Jeong, et al., 2002). This gene appears as a positive prognostic factor in our analysis. FSCN1 has been demonstrated to be a prognostic marker of invasiveness in Stage I NSCLC (Pelosi, et al., 2003), and appears as a negative prognostic factor in our analysis.


Some genes have been linked to other cancers. While it is possible that the connections between genes and lung cancer are circumstantial, we mention them here because some may be interesting and may turn out to be relevant to lung cancer. BCL9 is over-expressed in some cases of ALL (Katoh and Katoh, 2003), 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 (Kuo, et al., 2003). ATIC is a fusion partner of ALK that defines a subtype of anaplastic large cell lymphoma (ALCL) (Cheuk and Chan, 2001), 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 (Cairns and Walls, 1996).

None of the genes we have identified appeared in the list of top 100 genes from the Michigan analysis (Beer, et al, 2002), and we have only found one (CPE) that was mentioned in the Harvard paper (Bhattacharjee, et al, 2001). CPE was one of the genes defining a neuroendocrine cluster that they identified and associated with poor prognosis.

If our analysis is repeated separately for the Harvard and Michigan data sets, i.e., without pooling, only 8 and 1 of the 26 genes, respectively, are
flagged as having p-values less than 0.0024, while 17 are not flagged, including the top gene in our list (FCGRT). It is clear that we have obtained significant gains by pooling information across the two studies.

6. DISCUSSION

It may seem curious that our list of prognostic genes has almost no overlap with the genes mentioned in other publications based on these data, but this is reasonable for several reasons. First, we addressed a different research question than the analyses done in those publications. We used multivariable Cox models to search for genes offering prognostic information above and beyond what has been provided by known clinical predictors. In the study of Beer, et al. (2002), researchers looked for prognostic genes, but they fitted single-factor Cox models containing the gene expressions, but not clinical predictors. Thus, they were effectively searching for genes that provided information on survival, irrespective of whether the prognostic value of the gene was due to a possible association with known clinical factors like disease stage. Bhattacharjee, et al. (2001) approached the survival question indirectly by performing unsupervised clustering on the samples, testing which clusters had survival differences, then identifying the genes that were driving the clustering. Second, and perhaps more importantly, we gained increased power to detect prognostic genes as a result of pooling the data from the two studies.

There are clear benefits to be reaped by pooling information across microarray studies. Most microarray studies have small to moderate sample sizes, which means a relatively low statistical power that translates into a limited ability to detect significant relationships between gene expression levels and outcomes of interest. By pooling information across data sets, we can obtain additional sensitivity and specificity in identifying important genes. This may allow us to identify gene-outcome relationships that are undetectable in any one study alone. Of the 26 prognostic genes found in our analysis, 17 of them would not have been flagged by analogous methods in either the Harvard or the Michigan data without pooling. Given that many researchers make their data publicly available after publication, this suggests exciting possibilities for pooled analyses of existing data that could reveal important new insights into cancer biology.

Note that combining data across studies as we have done is fundamentally different from the pooling of results across studies that is typical in many meta-analyses. Pooling the actual data results in an increase
in statistical power to detect differences, while simply pooling the results does not. However, one must be careful in combining data across studies. First, one must account for any study-to-study heterogeneity that may be caused by differences in the studies’ patient populations or conditions. In this work, we dealt with this by incorporating a fixed effect for study in our survival models. If there are more than two studies available to pool, we recommend using either Bayesian hierarchical models (see Stangl, 1996) or frailty models (see Therneau and Grambsch, 2000), which both treat the study effect as random instead of fixed. These methods may not be as effective when pooling just two studies because they involve estimation of a variance component from a sample of size two.

Second, one must normalize the measurements to make them comparable across studies. In our case, this involves finding a way to effectively combine information across different microarray platforms. Here we have presented a new method that is applicable to oligonucleotide arrays in which we identify probes present on both platforms then combine them into new probesets based on Unigene clusters. Our investigations suggest that this method is reliable and precise, and yields comparable gene expression quantifications across two different versions of Affymetrix chips, the HuGeneFL and HG-U95Av2, used in the Michigan and the Harvard studies. We expect that this method may perform even better in combining information across U95 and U133 chips, since these chips have more probes in common. We feel that this approach is stronger than simply trying to normalize the expression levels across chips using quantile normalization, for example, since it is actually extracting measurements from the arrays that have scientific reasons to be comparable, and not just trying to make an arbitrary adjustment on non-comparable measurements.

Our specific biological goal in this analysis was to identify prognostic genes, meaning genes that offer information on patient survival beyond what is provided by known clinical predictors. We accomplished this by fitting multivariable Cox models that contained the clinical predictors along with the genes. It is important to adjust for these factors, since a gene that is simply a surrogate for a known clinical predictor is not as useful to us since we can gain the prognostic information directly from the clinical predictor without the additional time and expense required to collect microarray data. While this type of multivariable analysis may result in fewer flagged prognostic genes, we feel that this list has the potential to be more interesting biologically because we know that the flagged genes explain variability in patient survival not already explained by the clinical predictors. Many genes in our short list seem biologically interesting and have been linked with lung cancer in the existing literature.
1. **POOLING INFORMATION ACROSS DIFFERENT STUDIES**

AND OLIGONUCLEOTIDE CHIP TYPES TO IDENTIFY

PROGNOSTIC GENES FOR LUNG CANCER

7. **CONCLUSIONS**

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 pooled 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. Many of these genes would not have been found without pooling, and a large proportion of them appear to be biologically interesting and are worthy of future investigation.

8. **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.

9. **REFERENCES**


