Contents

CHAPTER 1

Alternative Probeset Definitions for Combining Microarray Data Across Studies Using Different Versions of Affymetrix Oligonucleotide Arrays

Jeffrey S. Morris, Chunlei Wu, Kevin R. Coombes, Keith A. Baggerly,
Jing Wang, & Li Zhang
University of Texas MD Anderson Cancer Center
Houston, TX, USA

1.1 Introduction

Many published microarray studies have small to moderate sample sizes, and thus have low statistical power to detect significant relationships between gene expression levels and outcomes of interest. By pooling data across multiple studies, however, we can gain power, enabling us to detect new relationships. This type of pooling is complicated by the fact that gene expression measurements from different microarray platforms are not directly comparable.

In this chapter, we discuss two methods for combining information across different versions of Affymetrix oligonucleotide arrays. Each involves a new approach for combining probes on the array into probesets. The first approach involves identifying ”matching probes” present on both chips, and then assembling them into new probesets based on Unigene clusters. We demonstrate that this method yields comparable expression level quantifications across chips without sacrificing much precision or significantly altering the relative ordering of the samples. We applied this method to combine information across two lung cancer studies performed using the HuGeneFL and U95Av2 chips, revealing some genes related to patient survival. It appears that the gain in statistical power from the pooling was key to identifying many of these genes, since most were not found by equivalent analyses performed separately on the two data sets. We have found that this approach is not feasible for combining information across the U95Av2 and U133A chips, which share fewer probes in common. Our second method defines probesets as sets of probes matching the same full-length

mRNA transcripts in current genomic databases. We found this method yielded comparable expression levels across U95Av2 and U133A chip types, and had better correlation across chip types than Affymetrix's matching probeset definitions.

1.2 Combining Microarray Data across Studies and Platforms

In recent years, microarrays have been used extensively in biomedical research. This is evident from the fact that there are over 9000 articles published since 2000 that involve microarrays, with over 3000 published in 2004 alone (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed). Generally, these studies involve the identification of individual genes or sets of genes whose expression profiles are related to clinical or biological factors of interest, including tissue type, disease status, disease subtype, patient prognosis, and biological pathway, to list a few. While microarrays measure the expression levels for thousands of genes, because of cost limitations, most studies are performed using only a small number of samples. As a result, individual studies often have limited power for detecting relevant biological relationships.

More recently, there has been a movement within the scientific community to make data from microarray studies publically available. This movement has been propelled by the establishment of standards for minimal information to provide when posting data (MIAME, Brazma, et al. 2001) and the requirement of many major journals to make such data publically available. There are currently a number of public repositories in which microarray data are posted, including ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/). This explosion of publically-available data makes it possible to consider meta-analyses that combine information across multiple studies, which allow one to assess the reliability of results reported in the individual studies and also to uncover new biological insights not discovered in any individual study. If done properly, this pooling of information across studies can provide increased power to detect small consistent relationships that may have gone undetected in the individual analyses, and can provide results that are more likely to prove reproducible.

There is a small but growing number of studies in existing literature that attempt to combine information across multiple data sets. Generally, there are three approaches that are used: 1. Identify an intersection of genes that are significant across multiple studies, 2. Validate results from a single individual study using data from other studies, or 3. Perform a single analysis after combining data across multiple studies. We now briefly discuss the merits and drawbacks of each approach.

The idea behind the first approach is that if a gene is truly differentially expressed, then this differential expression should be manifest across multiple data sets. However, this Venn diagram-based approach often reveals a shockingly small number of genes that are found to be differentially expressed in multiple data sets. In a study comparing normal and CLL B-cells, Wang et al. (2004) found that only 9 genes were

found to be differentially expressed in all three studies conducted on three different microarray platforms, out of 1172 that were differentially expressed in at least one study. Similarly, in a study involving pancreatic cells, Tan et al. (2003) found only 4 genes differentially expressed across 3 different platforms, among the 185 deemed differentially expressed on at least one platform. While perhaps identifying the most reliably differentially expressed genes, this approach actually results in reduced sensitivity for detecting biological relationships, since each (perhaps underpowered) study must find the gene significant before it is declared so. Other less conservative approaches focused on identifying genes that are consistent across studies include methods discussed in Rhodes et al. (2002) and Rhodes et al. (2004), which involve combining p-values across studies, and the integrative correlation method of Parmigiani, et al. (2004), which involves computing gene-gene pairwise correlations on the expression levels and/or tests statistics for each individual study, then computing a ”correlation of correlations” across studies. This approach results in a list of reproducible genes whose absolute or relative expression levels are correlated across studies and platforms. It does not, however, provide additional power for detecting biological relationships.

A number of studies take the second approach, identifying biological relationships using the data from a single study, then using data from other studies for validation of these relationships (Beer et al. 2002, Sorlie et al. 2003, Stec et al. 2005, and Wright et al. 2003). Since the studies may differ with respect to their patient populations, microarray platforms, and sample handling and processing, results surviving this stringent form of validation are likely to be real. However, like the first approach, this use of multiple data sets does not yield any additional power for detecting biological relationships since only a single data set is used in the discovery process.

In the third approach, the data is actually combined across studies and a single analysis is performed on the pooled data set. This is our primary interest in this chapter. The clear advantage of this approach is the possibility of increased power for detecting biological relationships, since the pooled data set is significantly larger than any of the individual data sets. The difficulty is that there are important differences between the studies that must be taken into account before it is possible to successfully pool the data. The studies may differ with respect to their patient populations, sample handling, or sample preparations. These differences can be manifest in both the clinical outcomes and the microarray data, and may affect the genes in a differential manner. It has been shown that it is possible to obtain comparable microarray data from different laboratories on a common platform if rigorous experimental protocols are established and followed across the different sites (Dobbin et al. 2005). However, posted data from different studies were likely generated using different protocols, so these factors come into play in the meta-analysis context. These problems are further exacerbated if the studies are conducted on different microarray platforms, which have technical differences that make their gene expression levels fundamentally incomparable (Kuo et al. 2002, Tan et al. 2003, Mah et al. 2004, Marshall 2004, Mechem et al. 2004a).

Some of this heterogeneity can be handled by modeling study effects for each gene

using fixed or random effects in the context of mixed models or Bayesian hierarchical models, standard approaches used in meta-analysis (Normand 1999, Ghosh 2004, Wang et al. 2004). These approaches appropriately account for the study-to-study variability when performing inference in the meta-analysis, and provide a simple first-order correction for each gene that aligns the mean expression levels for the different studies. Other approaches involve first-order corrections, but use methods that are more sophisticated mathematically. One is based on the singular value decomposition (Alter, Brown and Botstein 2000, Nielsen et al. 2002), and normalizes the raw expression levels within studies using the first eigenvectors for the genes and arrays. This approach assumes that these eigenvectors represent the study-to-study variability, which is assumed to dominate all other factors. Another approach (Benito, et al. 2004) normalizes using a new method called ”distance weighted discrimination” (DWD), which performs supervised discrimination to identify linear combinations of genes associated with the study effect, which is subsequently removed. However, these approaches, when applied to the raw expression levels, do not appear to be sufficient to make data comparable across different platforms. For one, they only adjust the mean of the distributions for the two studies, but do not adjust for higher order distributional properties like the variances or quantiles. In a study comparing data from spotted cDNA glass arrays and Affymetrix oligonucleotide arrays, Kuo et al. (2002) concluded that ”data from spotted cDNA microarrays could not be directly combined with data from synthesized oligonucleotide arrays”, and further, that it is unlikely that the data could be normalized using a common standardizing index.

For this reason, many studies do not attempt to combine the raw expression profiles across platforms, but instead only combine unitless summary measures derived from the raw data. The assumption is that, while the raw expression levels for the different studies may not be comparable, these unitless statistics should be, since they are at least on a common scale. For example, Wang et al. (2004) and Choi (2003) first compute the standardardized log fold changes between two experimental conditions, then combine these across studies using hierarchical models. Similarly, Ghosh et al. (2003) and Tan et al. (2003) first compute t-statistics comparing two experimental conditions, then combine these t-statistics across studies. Shen, Ghosh, and Chinnaiyan (2004) combine the posterior probabilities of being over-expressed, under-expressed, or similarly expressed between two experimental conditions across data sets. These approaches are promising and all result in increased power to detect biological relationships in the data, and can in principle be used across different platforms. However, we believe it would be inherently better to work with the raw expression levels, if we could get them to be comparable. In that case, we would not be limited to dichotomous comparisons, but could relate gene expression levels with any type of outcome (e.g. survival or time to progression). Also, these summary measures make implicit assumptions about the comparability of the reference populations in the different studies that, if not true, may adversely affect inference. For example, using t-statistics assumes that the mean and standard deviation of the true gene expression levels should be the same across studies, and are only different because of technical reasons. By using the raw expression levels, one could avoid making such assumptions.

Some studies have explicitly used sequence information to try to obtain comparable expression levels across platforms (Morris et al. 2005, Mecham et al. 2004a, Mah et al. 2004, Wu et al. 2005, Ji et al. 2005). This idea is natural, since much of the systematic variability between expression level measurements between (and even within) platforms is attributable to sequence-related factors, such as cross-hybridization, alternative splicing, inaccurate annotation of gene sequences, and RNA degradation. Cross-hybridization occurs when a gene hybridizes to ”near matches” on the array, which can attenuate estimates of gene expression. Certain sequences are more likely to cross-hybridize (Zhang et al. 2003), so may result in less reliable measurements of gene expression. Also, single genes may be transcribed into multiple different mRNA variants. These alternatively spliced variants may cause some sequences corresponding to different exons from the same gene to be discordant. Additionally, not all probes on microarrays map to annotated sequences in public databases. These probes tend to be less reliable (Mecham et al. 2004b), which may explain some of the lack of concordance across platforms. In a study involving matched samples run on Affymetrix and nylon cDNA arrays, Ji et al. (2005) showed that the correlation of expression levels these platforms was greater for sequences with matches in the RefSeq database. Finally, RNA degradation can affect probes differentially, since sequences closer to the endpoints of the gene may be more susceptible to this degradation than sequences near the middle. These factors are relevant when comparing completely different technologies, e.g. spotted glass cDNA arrays and Affymetrix oligonucleotide arrays, as well as when comparing different versions of the same technologies, e.g. different versions of Affymetrix arrays or glass cDNA arrays constructed using different clones. We believe that methods that explicitly take into account these known biological and technological factors ultimately will result in the most successful methods for combining information across platforms.

1.3 Overview of Affymetrix Oligonucleotide Arrays

Generally speaking, there are two major types of microarrays, cDNA arrays and oligonucleotide arrays. One key difference between these technologies is that on cDNA arrays, genes are represented by a single cDNA clone spotted on the array, while on oligonucleotide arrays (Lockhart et al. 1996), genes are represented by ”probes”, or short sequences of nucleotides from the target gene sequence. Affymetrix, Inc. (Santa Clara, CA) is the largest producer of oligonucleotide arrays, which they call GeneChips. Affymetrix GeneChips contain multiple probes for each gene. For the remainder of this chapter, we focus our attention on Affymetrix oligonucelotide arrays, which in practice are the most commonly used arrays today.

The Affymetrix probes each consist of a sequence of 25 bases from the target gene, which generally contains a total of several hundred or thousand base pairs. Since not all sequences bind equally well, there is natural variability between the expression level measurements for different probes taken from the same gene. In order to average over some of this variability, each gene is represented by a number of probes, which together form a ”probeset.” These probes are scattered across the array. For

each probe, there is also a corresponding ”mismatch” probe, which contains the identical sequence except with the 13th base replaced by its Watson-Crick complement. The mismatch probes are intended for normalization, although they have not been shown to be clearly useful for that purpose (Pope et al. 2004).

The probes are constructed based on sequence information contained in GenBank (http://www.psc.edu/general/software/packages/genbank/genbank.html), a public archive of DNA sequence information, Unigene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene), which partititions these sequences into non-redundant clusters presumably corresponding to genes, and RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/), which is constructed by the NCBI to represent the state of the art in terms of the sequences of known genes. As this information has evolved over time, Affymetrix has produced different versions of its GeneChip. The most commonly used chip types used in human studies include the HuGeneFL, the U95Av2, and the U133A.

The HuGeneFL was introduced in November 1998, and its sequence clusters are based upon Unigene build 18. It contains information on roughly 5600 genes, and each gene is represented by roughly 20 probe pairs. The probes corresponding to the same probeset are placed together in the same region of the array. The U95Av2 was introduced in April 2000, and is based upon Unigene build 95. It contains information on roughly 10,000 genes, each of which is represented by 16 probe pairs. The probes are randomly distributed across the array. The U133A was first introduced in January 2002, and is based upon Unigene build 133. It contains information on 14,500 genes, and contains 11 probes per gene. The probes are arranged on the array in such a way as to optimize the probe synthesis efficiency.

Frequently, researchers wish to combine information across experiments conducted using different versions of Affymetrix GeneChips. As new studies are conducted using more recent versions of the chips, researchers want to still use information from previous studies performed using older generations. Also, some researchers may want to perform meta-analyses on data collected from multiple studies performed at different institutions. It is not easy to merge information across chip types, since there are some genes represented on newer chips that were not on previous ones, and even the common genes are represented by different sets of probes on the different chips, so their expression levels are not generally comparable.

In the remainder of this chapter, we describe in detail two methods we have developed (Morris et al. 2005, Wu et al. 2005) to combine information across studies using different Affymetrix chip types. These methods use sequence information to define new probesets that yield comparable expression levels across different chip types. Our hope is that the raw expression level values using these redefined probesets are sufficiently comparable that they can be combined across versions. For each method, we describe the method and use an example data set to demonstrate the concordance of expression levels across different array types.

1.4 Partial Probesets

The incompatibility of expression levels across chip types is largely due to the fact that different sets of probes are used to represent the same genes on different chips. We expect, however, that individual probes present on multiple chips should yield comparable expression levels across chips. Thus, one approach for obtaining comparable expression levels across studies using two different chip types is to only use ”matching probes” that are present on both chip types.

For example, suppose we have microarray data from two studies, one performed on the HuGeneFL chip and the other on the U95Av2. The HuGeneFL contains a total of roughly 130,000 probes partitioned into 6,633 probesets, each containing 20 probe pairs, while the U95Av2 contains a total of roughly 200,000 probes partitioned into 12,625 probesets, each containing 16 probe pairs. There are a total of 34,428 ”matching probes” that are present on both chip types.

After identifying these matching probes, we then recombined these into new probesets based on the most current build of Unigene. We refer to these new probesets as ”partial probesets”. Note that because they are explicitly based on Unigene clusters, these probesets will not precisely correspond to Affymetrix-determined probesets. Frequently, multiple Affymetrix probesets map to the same Unigene cluster. We then eliminated any probesets containing just one or two probes, since we expected the gene expression measurements based on so few probes to be less reliable. When performed based on Unigene build 160, this left us with 4,101 partial probesets. In general, we expect these probesets to be smaller than the Affymetrix-defined probesets, since they only use the matching probes. Figure 1.1 contains a plot of the number of probes within each of these partial probesets. Most of the probesets (84%) contained 10 or fewer probes, and the median probeset size was seven. There were several probesets containing more than 20 probes.

1.5 Example: CAMDA 2003 Lung Cancer Data

Two independent studies were performed at Harvard University (Bhattacharjee et al. 2001) and Michigan University (Beer et al. 2002), both focusing on the same question of relating gene expression data to survival in lung cancer patients. These data were part of the 2003 critial assesssment of microarray data analysis (CAMDA) competition (http:/www.camda.duke.edu/camda2003). These studies both used Affymetrix GeneChips, but the Michigan study used the HuGeneFL while the Harvard study used the U95Av2. Our goal in analyzing these data was to combine information across both data sets to identify prognostic genes, whose expression levels provided prognostic information on patient survival over and above what is already provided by known clinical factors. We used partial probesets to quantify the gene expression levels, and demonstrated that this resulted in comparable expression levels across the two chip types, without any loss of precision from using only a subset of the probes. We identified a number of prognostic genes in our pooled analysis that were not discovered in the analyses performed on the individual studies, highlighting the benefit

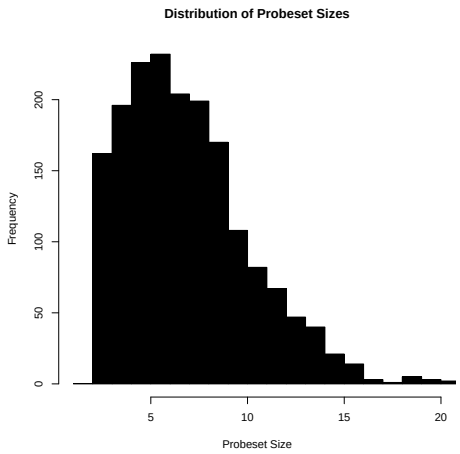

Figure 1.1 *Histogram of number of probes in each "partial probeset".*

of pooling data across studies. We first summarize these data sets, then describe our analyses to validate the partial probeset method and obtain prognostic genes. More details of this analysis can be found in Morris et al. (2005).

1.5.1 Overview of Data Sets

The Harvard study analyzed 186 lung tumor samples using U95Av2 Affymetrix GeneChips. From these, 125 were adenocarcinomas for which clinical information on the corresponding patients was available, including gender, age, stage of disease, and survival time. Applying hierarchical clustering to these data, Bhattacharjee et al. (2001) identified four distinct subtypes of adenocarcinoma with different molecular profiles, and further demonstrated that these subtypes had different survival prognoses.

The Michigan study analyzed 86 lung adenocarcinoma samples using HuGeneFL Affymetrix GeneChips. All of these samples also had corresponding clinical information, including gender, age, stage of disease, and survival time. Using univariate Cox regressions, they identified a number of genes whose expression levels were associated with patient survival. They subsequently constructed a "risk index" using the top 50 genes, and demonstrated that this risk index helped predict patient survival both in their own data and in independently obtained data from another experiment (Bhattacharjee et al. 2001).

In our own analysis, we first performed various quality control checks, after which we removed 10 arrays from the Michigan study and one from the Harvard study that demonstrated poor quality. This left us with a total of 200 arrays, 124 from the Harvard study and 76 from the Michigan study. Using the partial probeset definitions

described above, we quantified the gene expression levels for each partial probeset using the Positional Dependent Nearest Neighbor (PDNN) model (Zhang et al. 2003). Other quantification methods could have been used, but we chose this one because we believe its use of probe sequence information to predict patterns of specific and nonspecific hybridization intensities can lead to more reliable and accurate quantifications.

We also performed other preprocessing steps. We removed the half of the probesets with the lowest mean expression levels across all samples, then normalized 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 removed the probesets with the smallest variability across chips (standard deviation < 0.20), since we considered them unlikely to be discriminatory and more likely to be spuriously flagged as prognostic. Finally, we removed the probesets with poor relative agreement (Spearman correlation< 0.90) between the partial probeset and full probeset quantifications (see next section). After this preprocessing, 1036 probesets remained and were considered in our subsequent analyses.

1.5.2 Validation of Partial Probesets

Before analyzing the microarray data to identify prognostic genes, we assessed whether our method for combining information across different Affymetrix chip types performed acceptably. First, we checked whether the expression levels appeared to be comparable across chip types. Specifically, we computed the median and median absolute deviation (MAD) log expression level for each partial probeset across the Michigan samples run on the HuGeneFL chip and also for the Harvard samples run on the U95Av2 chip. Since the patient populations in the two studies appeared to reasonably similar, we expected to see high concordance in these quantities between the two chips if the expression levels were comparable. We did not, however, expect perfect concordance, since different patients were used in the two studies. Figure 1.2 contains a plot of these quantities, and demonstrates good concordance between the center and spread in the distribution of gene expression values on the two chips. The concordance between these values was 0.961 for the median and 0.820 for the MAD, so it appears that using the partial probeset method yielded reasonably comparable expression levels across the two chips.

Recall that partial probesets use only the matching probes, while completely ignoring expression level information for the non-matching probes. This means that partial probesets are generally smaller than the Affymetrix-defined probesets. The median size of our partial probesets was seven, 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 quantified the expression levels for the full probesets of the Harvard samples using the PDNN

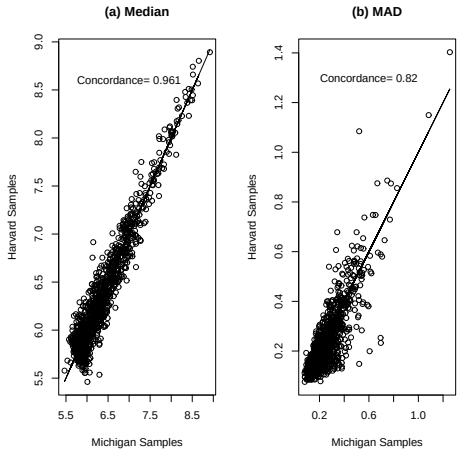

Figure 1.2 *Median (a) and median absolute deviation (b) expression levels for each partial probeset based on the Harvard samples run on the U95Av2 chips vs. the Michigan samples run on the HuGeneFL chip. The high concordance in these measures suggests we obtain reasonably comparable expression levels by using the matched probes.*

model. The full probesets consisted of all probes on the array mapping to the Unigene cluster, i.e., not just the matching ones. We plotted the standard deviation for each gene using the full probeset versus the standard deviation for the partial probeset, given in Figure 1.3. 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 was 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 retained in formulating the new chips may in some sense be the ”best” ones.

We computed 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 expected 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 was 0.95, suggesting that our method did 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. It appears that our partial probeset method worked quite well.

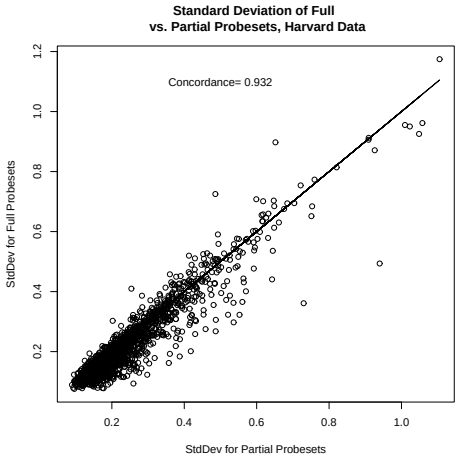

Figure 1.3 *Standard deviation across Harvard samples for each gene based on full and partial probesets. A "full probeset" contains all probes on the U95Av2 chip mapping to a unique Unigene ID, while the corresponding "partial probeset.*

1.5.3 Pooling Across Studies to Identify Prognostic Genes

We pooled the data across these two studies to identify prognostic genes offering predictive information on patient survival. We were not primarily interested in finding genes that were simply surrogates for known clinical prognostic factors like stage, since these factors are easily available without collecting microarray data. Rather, we were interested in finding genes that explained the variability in patient survival that remained 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 screened the 1036 genes to find potentially prognostic ones by fitting a series of multivariable Cox models containing age, stage (dichotomized as low, stages I-II, and high, stages III-IV), institution, and the log-expression of one of the genes as predictors. The institution effect was included in the model to account for differences in survival that were evident between the two studies, even after accounting for known clinical covariates. We obtained 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. We subsequently obtained the permutation p-value for each gene by counting the proportion of fitted Cox coefficients that were more extreme than the coefficient for the true dataset. A small p-value for a given gene indicated potential for that gene to provide prognostic information on survival beyond the clinical covariates. We also obtained p-values using asymptotic likelihood ratio tests (LRT) and the bootstrap to assess robustness of our results.

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, 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. We used this model to identify a p-value cutoff that controlled the false discovery rate (FDR, Benjamini and Hochberg, 1995) to be no more than 0.20. This means that of the genes flagged as prognostic, we expect at most 1 in 5 were false positives.

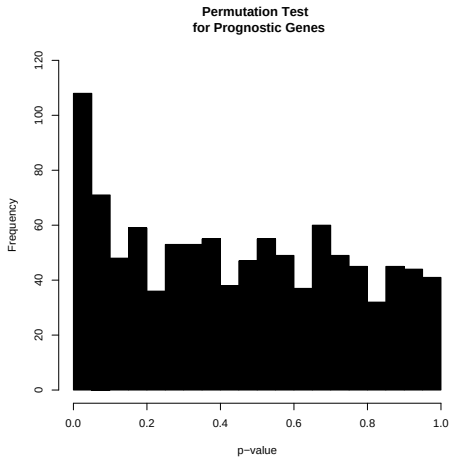

Figure 1.4 *Histogram of p-values from permutation test on gene coefficient in Cox model containing clinical covariates and each one of the 1036 candidate genes. The corresponding histogram for the LRT is nearly identical.*

Figure 1.4 contains the histogram of permutation test p-values. The overabundance of 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 1.1 contains a set of 26 genes that are flagged by the BUM method using FDR< 0.20, which are those genes with p-values less than 0.0025. Many of these genes appear to be biologically interesting and worthy of future consideration. We were able to link 10 of our 26 prognostic genes to lung cancer based on the existing literature. Four others could be linked to cancer in general or other lung disease in the literature. These genes are discussed in more detail in Morris et al. (2005).

None of the genes we identified appeared in the list of top 100 genes from the Michigan analysis (Beer, et al., 2002), and we 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. We repeated our analysis separately for the Harvard and Michigan data sets, i.e., without

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. A negative coefficient indicates that larger expression levels of that gene correspond to a better survival outcome.*

Gene Identity	Coef	Prognostic P-values		
		Permut.	LRT	Bootstrap
FCGRT	-2.07	< 0.00001*	0.00014*	0.0006*
ENO2	1.46	0.00001*	0.00002*	< 0.0001*
NFRKB	-2.81	0.00001*	0.00435	0.00404*
RRM1	1.81	0.00002*	0.00008*	< 0.0001*
TBCE	-2.35	0.00004*	0.00069*	0.0006*
Phosph. mutase 1	1.92	0.00008*	0.00020*	0.0004*
ATIC	1.81	0.00009*	0.00153*	0.0004*
CHKL	-1.43	0.00010*	0.02305	0.0260
DDX3	-2.37	0.00017*	0.00012*	0.0002*
OST	-1.64	0.00020*	0.00010*	0.0010*
CPE	0.72	0.00031*	0.00053*	0.0010*
ADRBK1	-2.20	0.00044*	0.00678	0.0030*
BCL9	-1.64	0.00067*	0.03602	0.0460
BZW1	1.33	0.00068*	0.00279*	0.0006*
TPS1	-0.64	0.00106*	0.00217*	< 0.0001*
CLU	-0.52	0.00109*	0.00239*	0.0024*
OGDH	-2.19	0.00118*	0.00405	0.0020*
STK25	2.29	0.00122*	0.00152*	0.0080
KCC2	-1.70	0.00143*	0.00988	0.0220
SEPW1	-1.29	0.00145*	0.01026	0.0160
FSCN1	0.66	0.00150*	0.00241*	0.0103
MRPL19	1.12	0.00211*	0.03213	0.0340
ALDH9	-1.18	0.00223*	0.00378*	0.0020*
PFN2	0.63	0.00248*	0.00351*	0.0020*
BTG2	-0.75	0.00232*	0.00580	0.0140

pooling, and only eight and one of the 26 genes, respectively, were flagged as having p-values less than 0.0025, while 17 are not flagged, including the top gene in our list (FCGRT). Thus, it appears that our pooled analysis revealed new biological insights contained in these data that were not identified when analyzing them separately.

1.6 Full-Length Transcript Based Probesets

The analyses presented in the previous section suggest that by using partial probesets, we were able to obtain comparable expression levels across studies conducted at different institutions using different chip types (HuGeneFL and U95Av2), allowing us to perform a pooled analysis that revealed new biological insights into lung cancer. Unfortunately, this approach is not feasible when combining information across the U95Av2 and U133A chips, since these chips share fewer probes in common than the HuGeneFl and U95Av2. There are 34,428 probes (14%) on the U95Av2 that are also present on the HuGeneFl, while there are only 11,582 probes (6%) that are also present on the U133A. If we form partial probesets and eliminate those with less than 3 probes, we are left with only 628 probesets. Thus, we have explored less stringent alternative approaches to use for combining information across these chip types.

One of the primary reasons probes yield discordant measurements is that they may be responding to different transcripts alternatively spliced from the same gene. When the transcripts are differentially regulated, the corresponding probes can yield conflicting signals. The current design of arrays ignores the effects of alternative splicing. Thus, if we differentiate the probes that match sets of alternatively spliced transcripts, we may be able to resolve the discordant measurements. Based on this idea, we developed a new method to regroup the probes into probesets. In our new definition of a probeset, all probes in the probeset must match the same set of full-length gene sequences. We refer to such a probeset as a ”Full-Length Transcript Based Probeset” (FLTBP, Wu et al. 2005). Assuming complete inclusion of alternatively spliced transcripts, we can in principle ensure concordant behavior of the probes within these probesets.

We now describe how we obtained these transcript-based probesets. First, we constructed a comprehensive library of full-length mRNA transcript sequences in the human genome by combining records in RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) and HinvDB (http:// hinvdb.ddbj.nig.ac.jp/index.jsp) databases. As of January 2005, RefSeq (build 111504, human section) contained 28,712 full-length transcript sequences representing 23,809 genes. H-InvDB (version 1.7) contained 41,118 sequences representing 21,037 genes. All of the sequences in this database were validated by full-length cDNA clones. We estimate that collectively the two databases represent approximately 29,000 genes with 50,000 non-redundant transcripts.

We used this library as the basis for defining our probesets. For each probe sequence used on the U133A and U95Av2 arrays, we identified all matching full-length transcripts using the Blast program (http://www.ncbi.nlm.nih.gov/blast/). We aggregated the IDs of those transcripts with exact matches to construct a matched target list. We found that 15% of the probes on the U95Av2 and 13% of the probes on the U133A had no exact match in our library, and 38% of the probes on the U133A and 33% of the probes on the U95Av2 matched more than two targets in our library, demonstrating that it was very common for one probe to match multiple targets.

By grouping the probes within the same matched target lists, we formed 23,972 and

14,148 probesets on the U133A and U95Av2, respectively. We call these probesets ”Full-Length Transcript Based Probesets” (FLTBPs). Because multiple probes in a probeset are essential to reduce noise and bias, we discarded all small probesets containing less than 3 probes, leaving us with 18,011 and 11,228 FLTBPs on the U133A and U95Av2, respectively. Collectively, these FLTBPs contained 82% of the probes on the arrays.

These new probesets were very different from the original ones. Only 9,893 of the original probesets on U133A and 5,257 original probesets on U95Av2 were the same after regrouping. Figure 1.5 shows a histogram of the number of probes in each FLTBPs. The probesets outside of the major peaks reflect division and fusion of the original probesets. Detailed information of our probesets are stored on our web site (http://odin.mdacc.tmc.edu/∼zhangli/FLTBP). This website also contains chip design files (CDF) using FLTBPs following the format designed by Affymetrix (http://www. affymetrix.com/index.affx). These CDF files can be used to run MAS5, RMA and dChip algorithms in Bioconductor (http://www.bioconductor.org/).

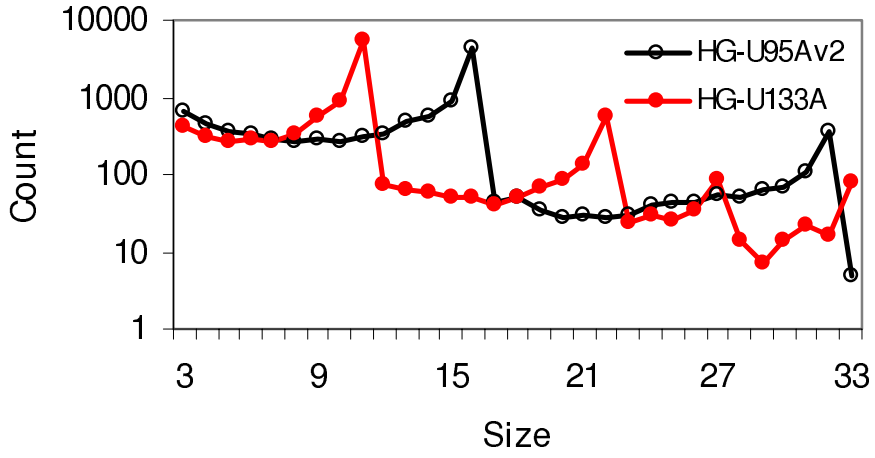

Figure 1.5 *Histogram of number of probes per FLTBP.*

By matching the matched target lists of FLTBPs on the two arrays, we found 9,642 pairs of FLTBPs that can be mapped between the U133A and U95Av2. Affymetrix has their own method for mapping probesets between different chip types (http://www. affymetrix.com/Auth/support/downloads/comparisons/best_match.zip), which yields 9,480 pairs of probesets between the U95Av2 and U133A chips. There are numerous differences between these Affy-defined mappings and our FLTBPs. Only 52% of the probe sets on the U133A and 48% of the probesets on the U95Av2 are mapped the same way as our FLTBPs.

1.7 Example: Lung Cell Line Data

To compare our mapping method with that of Affymetrix, we used a data set consisting of 28 paired measurements obtained by hybridizing identical samples on both the U133A and U95Av2 arrays. Because of this paired design, we expect very little biological variability between paired measurements on the two arrays, so any differences observed should be attributable to technical sources. We now describe this dataset and use it to demonstrate that the FLTBPs results in quantifications that are more comparable across chip types than Affymetrix- based probesets.

1.7.1 Overview of Data Set

Thirty RNA samples from variant lung cancer or normal lung cell lines and one human reference sample were hybridized on both U133A and U95Av2 arrays. Our quality control procedures revealed that three array images had obvious defects, so were discarded. This left us with 28 pairs of samples that we used in this study.

We preprocessed and quantified the gene expressions with PDNN (Zhang et al. 2003) using the PerfectMatch software (ver2.2) (http://odin.mdacc.tmc.edu/∼zhangli/ PerfectMatch). For comparison, we also preprocessed and quantified the data using other competing methods, RMA (Irizarry et al. 2003), MAS5 (http://www. affymetrix.com/ products/software/specific/mas.affx) and dChip (Li and Wong 2001), using bioconductor (v1.5, http://www.bioconductor.org/), following the default settings in the ”affy” package.

1.7.2 Validation of Transcript-Based Probesets

In order to assess comparability across chip types, for each gene, we computed the correlations between the paired U95Av2 and U133A measurements across samples. To enhance the contrast between two different mapping methods, in our comparisons we focused on the probesets that differed between the two methods. Approximately 1/3 of the probesets were mapped differently, which resulted in 3,309 and 3,527 paired probesets for FLTBP method and Affymetrix method, respectively.

Figure 1.6 contains a histogram of these correlations across probesets for the two mapping methods and four quantification methods. These histograms summarize the observed distribution of the paired correlations across probesets. Figure 1.6A clearly demonstrates that, when using the PDNN quantification method, the FLTBP mapping tends to yield better correlations than the Affymetrix mapping ($p < 0.00001$, Kolmogorov-Smirnov [KS] test). Notice the two peaks evident in the distribution of correlations for the Affymetrix mapping. The minor peak contains a large group of probesets with poor correlation across chip types. With other quantification methods, there is also evidence that the FLTBP method tends to result in better correlation across chip types than the Affymetrix method, although this evidence is not as strong

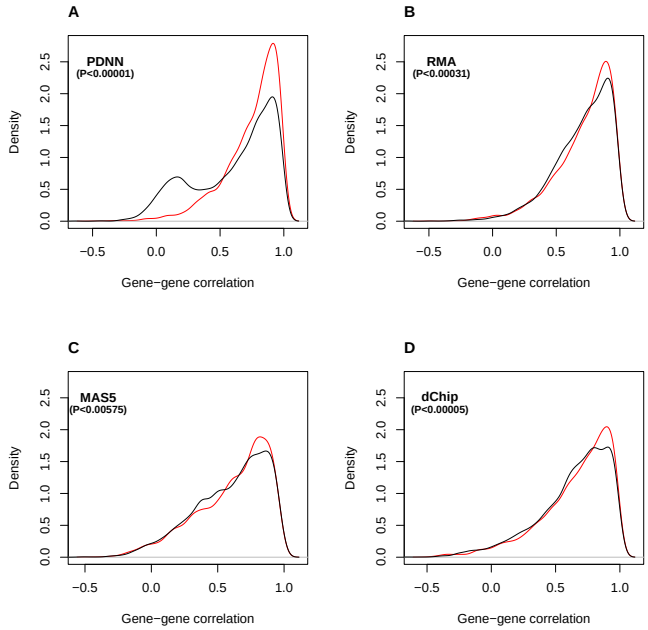

Figure 1.6 *Distribution of gene-to-gene correlation between probesets on two U95Av2 and U133A arrays, combining information over all samples, using both Affymetrix-defined probesets and FLTBPs. The correlations were computed using four different quantification methods, (A) PDNN, (B) RMA, (C) MAS5.0, and (D) dChip.*

(Figures 1.6B-D, $p = 0.00031$, 0.00575, and 0.00005 respectively). This improvement from using the FLTBPs is likely due to the fact that the FLTBP adjusts for some of the heterogeneity that is due to alternative splicing.

Note also that, when compared with Figure 1.6A, the distributions in Figure 1.6B-D are shifted more towards low correlations. This suggests that, for these data, the PDNN quantification tended to yield generally higher correlations than the RMA, MAS5, or dChip quantifications. This is even more evident in the sample-by-sample correlations between the chip types computed across genes, as shown in Figure 1.7. This increased correlation observed from the PDNN method may reflect the manner in which the PDNN model estimates and adjusts for the effects of non-specific binding.

From Figure 6A, we see that even when using the FLTBPs, not all genes displayed high correlations across chip types. Many of these low correlations were observed for genes that appeared to have low biological variability in these data. Low variability would make the noise component of the measurements dominate, resulting in low correlations. There are, however, some probesets with low correlations that do not have small variances. It is possible that some of the sequences corresponding to

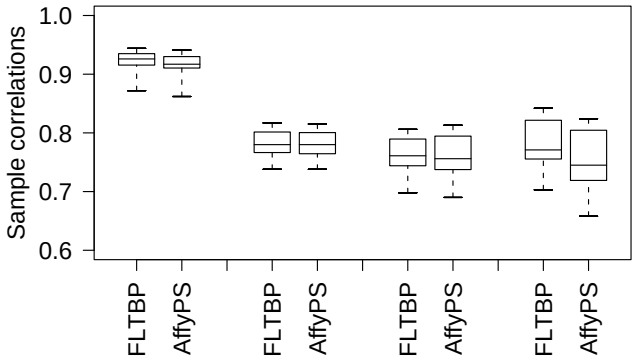

Figure 1.7 *Distribution of sample-to-sample correlation between probesets on two U95Av2 and U133A arrays, combining information over all genes, using both Affymetrix-defined probesets and FLTBPs. The correlations were computed using four different quantification methods, PDNN, RMA, MAS5.0, and dChip, respectively.*

these probesets were strongly affected by RNA degradation, or the currently available collection of transcripts may not include certain alternatively spliced variants that were differentially expressed across the sample tests, causing the correlations to become attenuated. Further work needs to be done to further reduce the effects of cross-hybridization and RNA degradation, which will hopefully lead to even more comparable expression levels across platforms.

1.8 Summary

In this chapter, we have illustrated the benefit of pooling data across multiple microarray studies. We performed a pooled analysis over two lung cancer microarray studies, and identified new prognostic genes that were not detected by separate analyses performed on the individual data sets. We also described two new probeset definitions that result in more comparable expression levels across different versions of Affymetrix oligonucleotide chips. The first method is based on partial probesets, which only use probes present on both chip types and combine them together based on Unigene cluster information. This approach works very well, but has limited applicability, since it is only feasible to apply across chip types that share many probes in common. The second method does not restrict us solely to matching probes, but works by recombining probes based on the set of full-length mRNA transcripts to which they map. In this way, the probesets map to the same set of alternatively spliced transcripts. Combined with the PDNN quantification method which accounts for non-specific binding, this approach appears to result in more comparable expression levels across chip types than Affymetrix's matched probesets. The benefit of

this approach is that it does not restrict attention to matched probes, so can be widely applied to combine data across any chip types. It may even be possible to use this principle to match up oligonucleotide array data with cDNA data, although this remains to be seen.

1.9 References

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

Beer DG, Kardia SLR, Huang C-C, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JMG, Iannettoni MD, Orringer MB and Hanash S (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 9, 816-824.

Benjamini, Y, and Hochberg, Y. (1995). 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.

Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1), 105-114.

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. *Proceedings of the National Academy of Science, USA* 98(24), 13790-13795.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* 29(4), 373.

Choi JK, Yu U, Kim S, Yoo OJ (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1, i84-i90.

Dobbin KK, Beer DG, Meyerson M, Yeatman TJ, Gerald WL, Jacobson JW, Conley B, Buetow KH, Heiskanen M, Simon RM, Minna JD, Girard L, Misek DE, Taylor JM, Hanash S, Naoki K, Hayes DN, Ladd-Acosta C, Enkemann SA, Viale A, Giordano TJ. (2005). Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clinical Cancer Research*, 11, 565-572.

Ghosh D (2004). Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics*, 20(11): 1663-1669.

Ghosh D, Barette TR, Rhodes D, Chinnaiyan AM (2003). Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional and Integrative Genomics*, 3(4), 180-188.

Irizarry, RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* 31(4), e15

Ji Y, Coombes KR, Zhang J, Wen S, Mitchell J, Pusztai L, Symmans F, and Wang J (2005). RefSeq refinements of UniGene-based gene matching improves the correlation between microarray platforms. *MD Anderson Department of Biostatistics and Applied Mathematics Technical Report*.

Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L and Kohane IS (2002) Analysis of matched

mRNA measurements from two different microarray technologies. *Bioinformatics*, 18, 405-412.

Li C and Wong WH (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proceedings of the National Academy of Science*, 98, 31-36.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13): 1675-1680.

Mah N, Thelin A, Lu T, Nikolaus S, Kuhbacher T, Gurbuz Y, Eickhoff H, Kloppel G, Lehrach H, Mellgard B, Costello CM, Schreiber S (2004). A comparison of oligonucleotide and cDNA-based microarray systems. *Physiological Genomics*, 16(3), 361-370.

Marshall E (2004). Getting the noise out of gene arrays. *Science*, 306, 630-631.

Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z (2004a). Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Research*, 32(9), e74.

Mecham BH, Wetmore DZ, Szallasi Z, Sadovsky Y, Kohane I, Mariani TJ (2004b). Increased measurement accuracy for sequence-verified microarray probes. *Physiological Genomics*, 18(3), 308-315.

Morris JS, Yin G, Baggerly KA, Wu C, and Zhang L (2005). Pooling Information Across Different Studies and Oligonucleotide Microarray Chip Types to Identify Prognostic Genes for Lung Cancer. *Methods of Microarray Data Analysis IV*, eds. JS Shoemaker and SM Lin, pp. 51-66, New York: Springer-Verlag.

Nielsen,T.O., West,R.B., Linn,S.C., Alter,O., Knowling,M.A., OConnell,J.X., Zhu,S., Fero,M., Sherlock,G., Pollack,J.R. et al. (2002) Molecular characterization of soft tissue tumours: a gene expression study. *Lancet*, 359, 1301-1307.

Normand SL (1999). Meta-analysis: formulating, evaluating, combining and reporting. *Statistics in Medicine*, 18, 321-359.

Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E. (2004b). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical Cancer Research*, 10(9), 2922-2927.

Pounds, S and Morris, S. (2003). 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.

Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, 62(15), 4427-4433.

Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Science USA*, 101(25), 9309-9314.

Shen R, Ghosh D, Chinnaiyan AM. (2004). Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*, 5(1), 94.

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. *Proceedings of the National Academy of Sciences*, 100(14), 8418-8423.

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

Stec J, Wang J, Coombes KR, Ayers M, Hoersch S, Gold DL, Ross JS, Hess KR, Tirrell S, Linette G, Hortobagyi GN, Symmans WF, and Pusztai L (2005). Comparison of the predictive accuracy of DNA array based multigene classifiers across cDNA arrays and Affymetrix GeneChips. *Journal of Molecular Diagnosis*, to appear.

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(19), 5676-5684.

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

Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV (2004). Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* 20(17), 3166-3178.

Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Science USA*, 100(19), 10585-10587.

Wu C, Morris JS, Baggerly KA, Coombes KR, and Zhang L (2005). A probe-to-transcripts mapping metehod for cross-platform comparisons of microarray data. *BEPress Technical Report.*

Zhang, L, Miles, MF, Aldape, KD. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology* 21(7), 818-821.
lasers", IEEE J Quantum Electron., Vol. 30, pp. 408–414, 1994.