Theses and Dissertations   Graduate School

2006

Design and Development of Oligonucleotide Microarrays and their Application in Diagnostic and Prognostic Estimation of Human Gliomas

G. Scott Taylor

Virginia Commonwealth University

www.manaraa.com

Design and Development of Oligonucleotide Microarrays and their Application in

Diagnostic and Prognostic Estimation of Human Gliomas

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy, Engineering, at Virginia Commonwealth University

By

G. Scott Taylor

B.S. Biology, Radford University 1997.

M.S. Biology, Virginia Commonwealth University 2002.

M.S. Engineering, Virginia Commonwealth University 2003.

Director: Kenneth J. Wynne

Professor, Department of Chemical Engineering

Virginia Commonwealth University

Richmond, Virginia

May, 2006

Acknowledgements

I would like to thank the VCU School of Engineering for the wonderful learning environment and excellent intellectual experience. I would also like to thank the members of my committee, Dr. Windle, Dr. Archer, Dr. Bowlin and Dr. Wynne for their helpful advise and patient councel. I also thank Dr. Peters and Dr. Overby for their sincere support. I extend my appreciation to my family and many collegues for all their help and mentorship. I convey my gratitude to Dr. Guiseppi for his mentorship and tireless work on behalf of the entire C3B research group. Finally, I wish to thank Colleen Higgins for her support and devotion.

Table of Contents

List of Tables

List of Figures

ABSTRACT

DESIGN AND DEVELOPMENT OF OLIGONUCLEOTIDE MICROARRAYS AND THEIR APPLICATION IN DIAGNOSTIC AND PROGNOSTIC ESTIMATION OF HUMAN GLIOMAS

By G. Scott Taylor

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University. Virginia Commonwealth University, 2005

Director: Kenneth J. Wynne

DNA microarrays represent an ultra-high throughput gene expression assay employed to study the transcriptomic profiles of biological tissues. These devices are increasingly being used to study many aspects of gene regulation, and there is growing interest in the biotechnology and pharmaceutical industries for developing such devices in efforts toward rational product/drug design. The DNA microarray also provides a unique and objective means for diagnosis and prognosis of human diseases based on patterns of gene expression. This is especially important in cancer research and the thrust toward personalized medicine. This dissertation details the design and development of

oligonucleotide microarrays and the design and execution of a gene expression study conducted using human glioma specimines. Chapter 2 details the design and development a ~10,000 gene human oligonucleotide microarray. This device consisted of a 21,168 features, each composed of a particular human gene-probe and was applied to the challenge of diagnostic and prognostic estimation for human gliomas (chapter 3). Gliomas are the most frequent and deadly neoplasms of the human brain characterized by a high misdiagnosis rate and low survival. The study in chapter 3 demonstrated that the specified design and development parameters were appropriate for conducting gene expression analysis and that this platform can be used successfully to predict malignancy grade and survival for glioma patients.

CHAPTER 1. MICROARRAY TECHNOLOGY AND CONTEMPORARY DATA ANALYSIS

ABSTRACT

Gliomas are the most frequent and deadly neoplasms of the human brain. Although most glioma specimines can be histopathologically classified with a high degree of accuracy, atypical gliomas are often difficult to classify by histological features and outcome prediction error prone. Indeed, within highly characteristic glioma specimens, much variation has been observed with regard to invasiveness, response to therapy, and ultimately prognosis. Recent advances in transcriptomic profiling have raised the possibility that DNA based devices can be developed to greatly improve diagnostic and prognostic information aiding the clinician in planning treatment and helping tailor treatments based on molecular characteristics of individual tumors. DNA microarrays represent the current state of the art with respect to high-throughput transcriptomic profiling. This chapter details the most important elements of microarray technology, glioma genetics, and introduces how our efforts at technology development have addressed critical issues with regard to design, fabrication, and molecular insight.

1

1.0 INTRODUCTION

The DNA microarray has been in use as a research technology for about ten years[1]. The initial platform consisted of poly-L-lysine coated standard glass microscope slides with cloned DNA (cDNA) polymerase chain reaction (PCR) products immobilized by ionic / electrostatic interaction with the surface. Microarray technology has evolved and now exists in three common platforms; *i*) photolithographically fabricated arrays, *ii*) the cDNA array and *iii*) the oligonucleotide array. DNA microarrays have further evolved to contain internal calibration and control features in addition to genomic probes (e.g., DNA probes obtained from the genome of an organism). The advantage the DNA microarray holds over older techniques is in throughput and the ability to assess correlative information to identify coregulated-gene networks (i.e., networks of genes whose regulation is dependent on the other network members). The potential of this technology is that all of the genes in the genome of an organism can be arrayed and immobilized on a consistent single substrate in specific locations on the micrometer scale. This ultimately enables an instantaneous snapshot of the entire active transcriptomic profile (i.e., the profile of messenger RNA (mRNA) transcripts present in the cell) in any given cell population or tissue. The huge advantage in throughput allows the investigation of not only gene-wise and sample-wise differences in expression patterns, but also biological complexity.

Microarrays have had a tremendous impact on cancer research allowing investigators to discern gene expression networks as revealed in comparative analyses of

tumor and normal tissue extracts and as diagnostic tools that essentially function by pattern recognition. It has been reported, for instance, that microarray data can be more accurate in predicting survival than histopathological grading for ambiguous samples[2]. Further, the concept that gene expression patterns can help define treatment options has been established widely [2-7], particularly lung cancer [8], colon[5], breast[9-11], and leukemia[7,12]. These advances provide impetus for the development of a new class of devices that utilize an empirically defined, targeted suite of genes to improve diagnostic resolution prognostic estimation, and maximize the effectiveness of treatment regimens.

A major challenge faced by contemporary medicine is diagnostic and prognostic estimation for brain cancer. Malignant glial tumors of the central nervous system, collectively referred to as gliomas, present one outstanding example of the need for improvement in predicitive technologies to inform treatment. Glial tumors are the most frequent and deadly human neoplasms of the brain and kill and estimated 13,000 -17,000 Americans per year[13]. Decades of research into the cellular and molecular biology of glial tumors (astrocytomas (AA) glioblastoma multiform (GM) and oligodendriogliomas (OL)), has revealed a more coherent picture of the biology of these deadly neoplasms. This effort has been aided and enhanced by the application of DNA microarrays, although patient survival has not improved in 25 years[13]. Nevertheless, such data has already yielded valuable insights into options for patient therapy. For example, the presence of DAP-3, a protein associated with cellular motility and radiation resistance[4], was reported as over expressed in the invasive rim of a characteristic glioma. This protein notably belongs to an anti-apoptotic network, suggesting that therapies aimed at inducing

apoptosis, such as the chemotherapeutic cis-platin, or radiation therapy (whole brain and gamma knife), are unlikely to eliminate the disease[4]. Indeed, there is only marginal survival benefit correlated with the application of current treatments including chemotherapy, radiotherapy, and surgical debulking of the tumor[14].

Currently, the two most important factors in brain tumor survival are age at diagnosis and tumor type histology (grade of malignancy). It is commonly believed that DNA based devices will be integral in improving patient care and treatment while simultaneously contributing to the identification of genetic targets for novel therapeutics. This chapter is devoted to presenting DNA microarray technology, experimental design, data analysis, and glioma biology and contemporary microarray classification efforts.

1.1 DNA MICROARRAY TECHNOLOGY

1.1.1 Preliminaries

DNA microarrays (MA) have emerged as a powerful technology for capturing the instantaneous profile of messenger RNA (mRNA) transcripts (transcriptome) within any given cell population[3,4,6,11,12,15-20]. A DNA microarray is composed of hundreds to thousands of individual genes (probes), immobilized in a grid of discrete spots. There are three commonly used platforms for DNA microarrays; *i)* in-situ photolithographically synthesized oligonucleotide microarrays (typified by Affymetrix and Protogene microarrays), *ii)* spotted cDNA microarrays, and *iii)* spotted oligonucleotide (oligo) microarrays[21], the latter two platforms are produced by non-contact or contact deposition of the nucleic acids.

Post fabrication, neglecting experimental design, microarray experiments are conducted in five basic process steps *i)* harvesting of messenger RNA (mRNA) from biological cells or tissue, *ii)* enzymatic replication and fluorescent labeling of harvested

Figure 1.1. Schematic of the microarray experimental process.   1. mRNA is harvested from biological cells. 2. Harvested mRNA is reverse transcribed and labeled. 3. This product is applied to the surface of the microarray for the hybridization reaction, 4. After hy   bridization the microarray is scanned. 5. The scanned image is quantified and the data is analyzed. (Figure from http://www.ambion.com/techlib/resources/microarray/basics1.html   )

mRNA, *iii)* hybridization of labeled target cDNA to immobilized probe on the array, and *iv)* scanning and *v)* data analysis. Each step in the process has an associated set of variables, and removing and/or standardizing variables in a process step is highly desirable as microarray data is strongly influenced by variation imposed by these variables.

Microarrays have major application in cancer research but also in expression analysis (drug discovery, development, toxicology), single nucleotide polymorphism

analysis[22], gene discovery[21], diagnostics[23-25], genetic sub-typing[10,15,26,27]. While this is not an exhaustive list, it illustrates the range of biological questions being addressed through microarray analysis.

Concomitant with development of microarray technology has been the investigation into methods for experimental design (ED), normalization, analysis, and mining the extremely large amounts of data generated. While there are many types of analysis that can be performed on microarray data, and finding an appropriate analysis protocol can be challenging, there are guidelines to consider that aid in establishing effective analysis. For instance, the nature of the query relates to the ED, which in turn influences statistical precision. A loop design[28] may be more effective for a drug *vs*. cell line interaction study since there are limited conditions and statistical precision is high for measures of significance. On the other hand, a reference design[28] may be more appropriate for a classification study where there could be hundreds of biological and technical replicates and clustering methods are used that would be convoluted by the complexity of a loop design.

1.1.2 Microarray Design and Fabrication

Microarray design takes into consideration the types and placement of features (DNA-probes) for such parameters as intensity calibration, normalization, grid alignment, positive control for hybridization, uniformity of hybridization, fidelity of reverse transcription, and replication. By definition an array of gene-probes (features) is a 2D arrangement of rows and columns of spots. This array is further partitioned into sub-arrays (subgrids) as a consequence of the utilization of multiple spotting pins during

fabrication. Subgrid rows and columns are often referred to as meta-rows and meta-columns. Control and calibration features are typically embedded within the array to gauge assay preformance.

Intensity calibration and normalization was traditionally performed through the inclusion of so-called housekeeping control genes, randomly dispersed, throughout the array[21,29, 30]. However, this method had its drawbacks as the genes originally considered to have stable expression were actually found to be quite variable. A somewhat more sophisticated method was the use of an "invariant set", a set of constantly expressed genes that could be determined during a pre-analysis step[31] Yang, et al., reported the inclusion of a 'microarray sample pool' feature composed of each probe present on the microarray and deposited in a dilution series, to provide normalization parameters for non-specific hybridization[32]. During image quantification, it is useful to have "land marks" placed on the microarray sub-grid corners, to aid in alignment of the segmentation grid. Such features can be genes known to have high abundance, or spiked-in control probes that bind to pre-manufactured, labeled, complimentary DNA spiked into the hybridization solution. Features such as these are also useful as positive controls for hybridization and uniformity of hybridization. They are typically represented by genes from an organism other than the one for which the bulk of the probes interrogate. Probes that are complimentary for the 3′ and 5′ ends of a gene can be used to test for fidelity of the reverse transcription and RNA integrity, the idea being that if reverse transcription was preformed efficiently and RNA integrity is acceptable, than the ratio of the intensity

from these spots should be close to one. Finally, it is useful to replicate each gene-probe on the microarray as this replication can improve sensitivity and statistical precision[32,33].

Fabrication of spotted DNA microarrays in its most general form is the process of spotting single stranded (ssDNA) or double stranded DNA (dsDNA) on to the surface of a particular substrate. This substrate historically has been nitrocellulose, nylon, or boro-silicate glass (microscope slide). The array is produced using high precision robotics to iteratively aspirate and dispense DNA fragments onto the substrate in a 2D grid arrangement. The spotting is done by simply dipping specialized stainless steel or silicon pins into a solution containing the DNA then contacting the substrate, thereby ejecting a tiny fraction of the aspirated DNA solution onto the surface of the substrate. Factors that influence the performance of the microarray include surface chemistry, spotting concentration, spotting buffer, type of printed DNA, type of printing pin, production time, length of production time, ambient humidity, production batch, and the curing process.

Custom spotted microarrays are typically produced using a surface modified standard microscope slide. Naturally, there are several surface modifications for microarrays which can be categorized with respect to their surface interaction with the DNA as covalent, non-covalent, or hydrogel. Poly-L-lysine (PLL), γ-aminopropyltrimethoxysilane (APS), and amino dendrimers, are examples of surface modifications that interact with DNA non-covalently (Figure 2.2). PLL was the first surface modification used for microarrays[34], and is still commonly used. Generally

surfaces that present free amines interact with DNA non-covalently. These surfaces are most useful when printing unmodified PCR products.

Epoxy-silane (3-glycidoxypropyltrimethoxysilane (GPS)) and aldehyde-silane surfaces are used in conjunction with 5'-NH$_2$-C$_6$ terminated oligos. A covalent bond is formed between the epoxide ring (or carboxylic acid) and the amine terminus of the oligonucleotide probe (Figure 2.2). These surfaces do not require a blocking step prior to hybridization with the sample and exhibit reduced background with respect to the amine surface[35]. The reason for this, we suggest, is a propensity for labeled cDNA to interact electrostatically with the amine surfaces, hence the perceived necessity for blocking. Notwithstanding, pre-hybridization blocking on GPS compared to amine surfaces is associated with lower background for the GPS surface (Chapter 2)

Spotting concentration for DNA microarrays has been studied in great detail. It has been reported that concentrations greater than ~6-10 µM do not significantly improve intensity[35,36]. Zammatteo, et at., reported a maximal density of the probe of 600 fmol/cm^2 which was reached a printing concentration of 0.5 µM[37]. We demonstrated that increasing the spotting concentration of oligonucleotides over four orders of magnitude (0.0001µg/µl –1µg/µl) resulted in only a marginal increase in signal intensity regardless of the surface chemistry[35] (Chapter 2).

1.1.3 Experimental Design for Microarrays

1.1.3.1 General Design Considerations and Sample Size

The practice of ED is traditionally coupled with the specific aims of the study while emphasizing the parsimonious use of resources. The specific objectives and study

design largely determine the statistical methods that will be used. When the purpose of the study is to determine which factor has the greatest influence on the response variable, analsis of variance (ANOVA) is the appropriate analytical method which models the response's dependence on important explanatory factors [38]. Experimental design is naturally, constrained by the goal of the study, the number of sample classes, desired statistical power, efficiency, sample availability, type and level of replication, and so on. However, the goal of ED is to maximize the information generated given the practical constraints. The EDs discussed in subsequent sections are limited to a two-dye sample labeling system.

The basic question for which microarray technology was developed, is finding differential gene expression patterns among interesting biological comparisons. This is known as differential gene expression analysis (DGEA). Approaches to answering this question depend on the goals of the study. Methods for statistical inference such as the student's t-test, ANOVA, significance analysis of microarrays (SAM), and proportional hazards (regression) all have been successfully employed for DGEA on a gene-wise basis[28 39-41]. In addition, such algorithms can also be used for feature selection for class prediction investigations or for pathway studies[28 39-41].

There are several ED's that are particularly useful for microarray studies. For a two-dye system, each spot on a MA can be considered an experimental unit with block size = 2. In this situation experiments with more than two comparisons are by default an incomplete block design, which may be balanced or unbalanced. Due to differences in spot uniformity, probe concentration, and hybridization uniformity, the value of

hybridizing two samples to an array, is that the spot performance is "controlled"for by the relative comparison. Microarray experiments must be executed in such a way that all comparisons of interest are estimable a concept that underlies the ED approach.

There is generally thought be a significant effect due to the labeling dye. Consider a situation where there are only two samples, one sample is labeled with dye 1, and the other with dye 2 and an MA is hybridized. A researcher might decide to control for the dye effect by switching the labeling assignments and performing another hybridization. This set up is known as a dye-swap design, Figure 1.2. This design is very efficient with respect to the estimation of statistical parameters, however it becomes impractical when
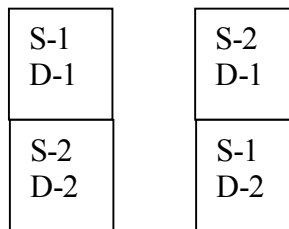
Figure 1.2. Dye-swap design. Samples are denoted S, dyes are denoted D, each rectangle is an array

there are more than two classes compared. Differential gene expression analysis is carried out by combining the per sample gene-wise intensity measures and fitting a one-way model such as in Eq.1.6.

For larger numbers of sample classes, an alternative design is the loop design. Here, each sample is labeled with each dye and are paired on arrays as depicted in Figure 1.3. This design is also more efficient than the reference design until sample types reach about ten[42]. This design is useful when there is only one factor (main effect). Investigations with greater numbers of factors (i.e.: two cell lines subjected to two different drugs) require another approach to ED.

Figure 1.3. Loop design for microarrays. Each arrow represents a microarray, the circular end denotes labeling with dye 1, and the pointed end represents labeling with dye 2.
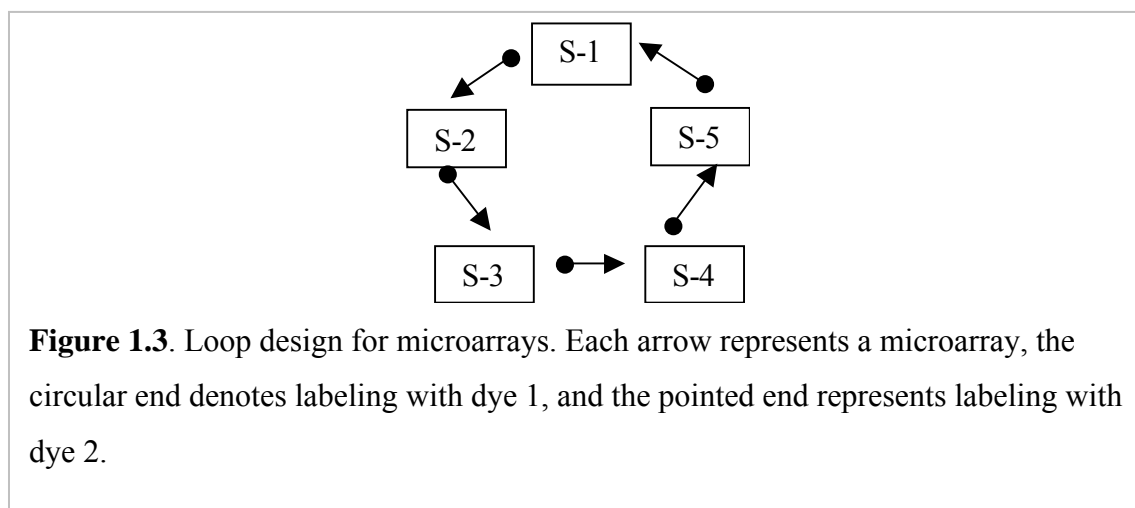
While a loop design may be more efficient and precise for differential gene expression analysis (DGEA), it is impractical for large sample sets. If one of the arrays is lost or performs poorly, the ability to perform comparsons breaks down. Further, if more samples are necessary in future analysis, it is difficult or impossible to integrate them. However, for small sample sets the, greater level of replication provides an advantage in terms of resource commitment over a replicated reference design[42,43].

The reference design, depicted in Figure 1.4 is performed by co-hybridizing a reference sample to each array along with the sample of interest. Because the reference sample is always labeled with the same dye, a dye-swap design is not necessary. In

addition the design as the advantages that it is simple, easily extensible, and facilitates down stream analysis such as clustering easier than a loop design. A disadvantage of this design is that for small numbers of samples, it is not as efficient as the dye-swap or loop design. This translated into reduced precision in estimating DGE.

The take home message of the above discussion is that because of the interplay between ED and the analysis procedures, researchers should carefully consider the down-stream implications of selecting a design and whether the resources will be available to properly utilize the advantages of the selected design.

Sample size (or level of replication), as a function of desired statistical power for microarray data, can be estimated *a-priori* from preliminary data. These quantities also depend on the precision of model parameter estimates, which is in turn a function of design efficiency. For instance, because a loop design is more efficient than a reference design, the level of replication in a loop design may be lower to achieve the same level of statistical power. Again, estimation of sample size should be considered in the context of the aim of the study.

One simple method for estimating sample size was described by Simon[44], et al., where statistical power, P = (1 - β), for a given value of β is indicated in Eq. 1.1,

$$n = 4\left(z_{\alpha/2} + z_{\beta}\right)^2 / \left(\delta/\sigma\right)^2 \qquad [1.1]$$

Here, n = number of samples, $z_{\alpha/2}$ = percentiles of standard normal distribution z_{β} = percentiles of standard normal distribution , δ = effect size σ = within class standard deviation, where σ is estimated from prior data. The values of α and β must be chosen with regard to multiple testing considerations. This is the general rule for specifying the

type I and II error rates for microarray data since there are so many individual tests being carried out during the analysis (equal to the number of genes or features on the array). Generally, researchers are more tolerant of a moderate value for β (0.05) than for α.

When the goal of the study is to develop prognostic models the following was also method reported by Simon, et al., to estimate sample size,

$$D = \left(z_{\alpha/2} + z_{\beta}\right)^2 / \left(\tau \ln \delta\right)^2 \qquad [1.2]$$

Where τ is the standard deviation of the gene-wise log ratio or intensity over all samples, δ denotes the hazard ratio related to a unit change in the log ratio[44].

This model takes into account that survival data are usually continuous and right censored, and that there are not discrete differences between sample types as in a cell line study.

Finally, technical replication in the form of replicated spots on the array and hybridization of the same sample on multiple arrays provides estimates for measurement error. Biological replication provides information on the distribution of gene expression values for a given gene among individuals in the population. Thus, the amount and type of replication are dependent on the objectives of the study. If the study is of the response of a cell line to a type of drug, then technical replication improves sensitivity. On the other hand, a study of tumors aimed at novel class discovery relies heavily on estimating the variability across as many individuals as possible. The following discussion addresses the characteristics, merits and demerits of EDs for microarrays.

1.1.3.2 The Reference Design

The simplest design choice is a direct comparison of two samples (e.g., tumor *vs.* normal) on one microarray. Experiments where there are more than two comparisons are most effectively understood as incomplete block designs where each array (or spot) constitutes a block of size *d* where *d* = number of dyes. If there more than two sample

Figure 1.4. A schematic of a reference design. Each box represents an array and each sample (S_i) is compared to the same reference RNA.

varieties (*V*), then we cannot assay all *V*s on a single array and varieties must be assayed in an alternative way to make all relevant comparisons possible. Let V_i be the variety of sample being hybridized, where i = 1… n and for each array two varieties are applied. In a reference design V_i is co-hybridized with a standardized reference V_r, and for differential expression between varieties we are, in general, testing H_o: $V_1 = V_2 = … = V_n$. Note that the values of *V* are the measures of expression used to calculate significance and are ratios: generally, $V_g = I_1/I_2$, where I is the estimated fluorescence intensity at wavelengths 1 and 2 respectively. Since each variety of interest is compared to the same reference the distance between any two samples is the same, which makes model fitting and subsequent analysis easy. This design also has an advantage when large numbers of samples are to be analyzed, it is extensible, and samples can be collected in a somewhat

haphazard fashion. However, due to the large degrees of freedom cost associated with the reference sample, it is less efficient and parameter estimates are less precise.

This design is perhaps the simplest to execute and has the advantage of being extensible such that samples can be assayed somewhat haphazardly[42,44]. Downstream analysis such as clustering is also easy because the distance between samples is always the same[45]. One criticism of this approach is that dye effects are completely confounded with sample effects due to the fact that the reference is always labeled with the same dye. This precludes specifying the DG term in an ANOVA model. However, in practice this is of no consequence since researchers are not usually interested in the reference channel, which essentially serves to correct for differences in feature performance. Another criticism is that the reference design is inefficient compared to other designs. For large sample sizes the reference is perhaps the most parsimonious choice since loop and dye-swap designs become impractically complex in execution and analysis.

In summary, by far, the reference design is the most widely used design for two-channel microarray hybridizations due to its simplicity of execution and intuitive nature. The statistical properties of the reference design are often overlooked by the researcher and can remain an afterthought due to its robustness. The biggest pitfall in utilizing this design is that it is inefficient and statistical inference can be imprecise for small effect sizes. One way to insulate against this is to design the microarray itself with replicate spots and include as many technical and biological replicates as possible.

1.2 DATA ANALYSIS FOR DNA MICROARRAYS

When considering analysis, it is useful to group the workflow into three categories: *i)* pre-treatment, *ii)* normalization, *iii)* down stream analysis (i.e.: gene expression analysis, clustering, and so on). During pre-treatment, raw expression data are typically subjected various operations such as background subtraction, Log_2 transformation, intensity filtering or variance filtering. Normalization methods are meant to remove systematic noise from the data and can be applied within an array or across arrays and/or both. Down stream analysis seeks to extract the biologically relevant information contained in microarray data. Process such as differential gene expression analysis[45-48], hierarchical clustering[41,42] supervised learning[7,26,39,49], are commonly performed to test hypotheses, discover gene interactions[50], discover novel tissue sub-types[2], delineate sample groupings[41], or predict class membership[5,6,11,12,16,17,23,25,45,47,51]. Since we are primarily interested in predicting class membership (i.e., histopathological class or survival) the following sections describe contemporary methods for DGEA and class prediction problems.

1.2.1 Data Normalization

Data *transformation* is the first step in the analysis process, and is applied to stabilize variance or rescale and remove distributional artifacts caused by systematic noise[21,36]. Data pre-treated in this way are then normalization by means of a number of algorithms, the choice is left to the discretion of the researcher. A common post-normalization step is filtering, which can be done by intensity, variance, coefficient of variation, or some other statistic. Control features (spots) embedded in the array can

enable subsequent normalization and calibration. Common applications for within array and between array normalization include total-intensity, invariant set, mean or median centering, ANOVA, to standard deviation regularization, and scatter plot smoothers (LOWESS)[21,28,32,45,52].

Normalization of microarray data has been the subject of a growing body of literature; as with other data treatments related to microarrays, there is little consensus, and the community seems to be taking a situation specific approach. Nevertheless, certain trends have become established. For instance, most available MA analysis software has an implementation of LOWESS normalization (Biodiscovery, GeneSpring, TMeV) and many MA studies report using this normalization during the analysis[26,32,53,54,55]. Further, total-intensity normalization has been criticized due to its potential for over smoothing[32,56], and normalization to so-called 'housekeeping genes' has been similarly debunked[29].

The LOWESS algorithm is probably the most widely used normalization and is essentially a scatter plot smoother developed for other applications[57]. It implicitly assumes that 95% of genes have no expression change. In practice, the rule of thumb is ~70%. This algorithm fits a locally weighted polynomial to a neighborhood of points, determines the weighted least squares fit on intensity measurement, s_{sg} and computes a fitted value $\hat{s}_o = w(x_o)$. The probe intensities are then adjusted by $x^{norm} = 2^{A+\S/2}$. This algorithm proceeds through the following steps, $i)$ on an MA plot, otherwise known as the ratio vs. intensity plot, transform the data such that $A = \frac{1}{2} \text{Log}_2(x*y)$ and $M = \text{Log}_2(x/y)$ where x denotes the cy5 intensity of gene (g) and y denotes the cy3 intensity

value of gene g, *ii)* take a point (x_o) and find *m* nearest neighbors according to a specified observational space *f,* typically $f \approx 0.4$. *iii)* Compute the Euclidean distance from x_o, $|x_i - x_o|$, *iv)* compute the largest distance between x_o and another point in the neighborhood, $\Delta(x_o) = \max(|x_i - x_o|)$. *v)* Assign weights to each point in $N(x_o)$ using the tricube weight function:

$$(|x_i - x_o| / \Delta(x_o) = u \text{ , and } w(u) = \begin{cases} (1-u^3)^3, 0 \le u \le 1 \\ 0 \qquad , otherwise \end{cases} \qquad [1.3]$$

which basically assigns a smaller weight to points further from x_o. *vi)* Calculate the weighted least squares fit on y_{sg} on the neighborhood $N(x_o)$, and take the fitted value using $w(u)$ as the weights. *vii)* Repeat this procedure for each xs_g and adjust the probe intensities by $x^{norm} = 2^{A+\hat{s}/2}$.

This type of normalization has been shown to improve gene expression measurements in self *vs.* self hybridizations for custom spotted microarray analysis[32], and can be applied with respect print-tip to take into account variations produced by the individual printing pins. LOWESS normalization has the overall effect of re-centering the distribution of gene expression values around zero on the y-axis of an MA plot and the lower the value of *f,* the more aggressive the smoothing becomes.

Normalization can be accomplished implicitly in an ANOVA setting by fitting a model such as Eq. 1.4 to the data.

$$y_{akghv} = \mu + A_a + D_k + G_g + AD_{ak} + VG_{ig} + DG_{kg} + AG_{ag} + S_{h(a)} + \varepsilon_{akghv} \qquad [1.4]$$

where a for m arrays, k for o dyes, g for p genes, h for q spots, i for v treatments/classes/groups.

Fitting this model to the data has the effect of partitioning the measured intensity y_{akghv} into various sums of squares and removing their quantities from the error term, thus making the *F*-ratio larger (and more significant). Here μ is the overall mean of all expression values for all arrays, A is the effect of array (a), D is the effect of dye (k), G is the effect of gene (g), V is the effect of variety (i), and S is the effect of spot (h). Combinations of terms, i.e.: AD_{ak}, denotes the interaction of the ath array with the kth dye.

This approach was initially described for microarrays by Kerr et al., and Wolfinger, et al., and was demonstrated to produce a robust normalization that takes into account array and dye effects and any other source of systematic noise that can be
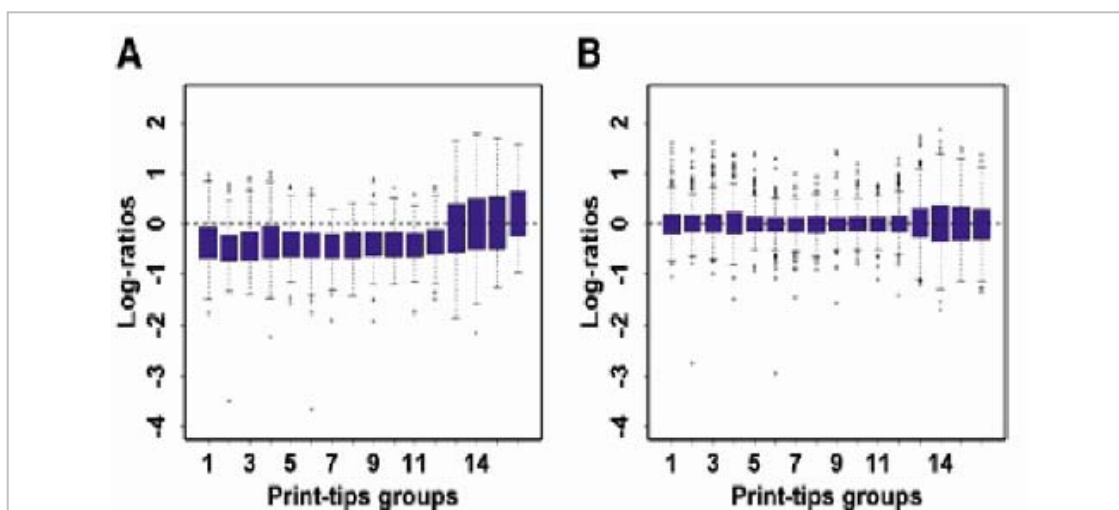
Figure 1.5. Box plots before and after print tip LOWESS normalization. A. Before LOWESS. B. After LOWESS.

encoded into the model. Examples include day of the week, experimenter, print tip, and

so on[28,43]. In practice, approach is taken in the context of statistical inference and

normalized data from the model fitting step is not typically available for other analyses

due to its limited software implementations.
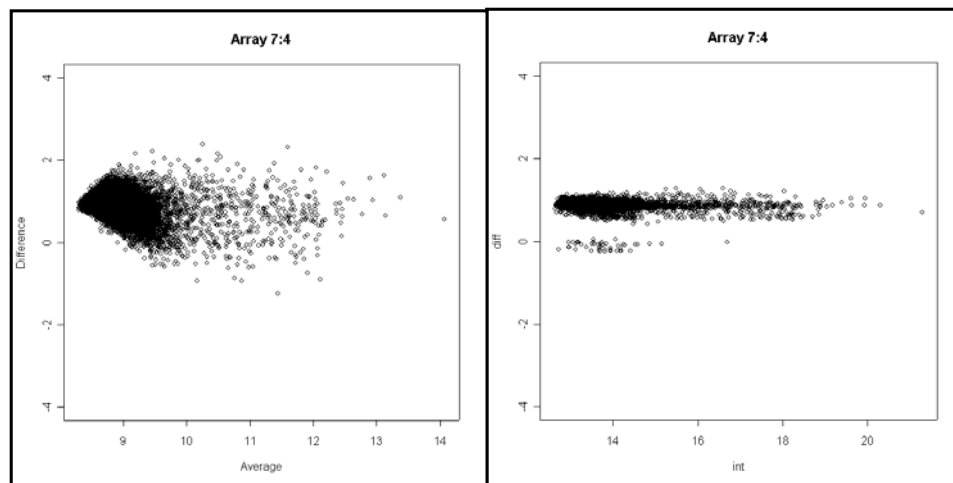
Figure 1.6. The effect of ANOVA normalization. Ratio *vs.* intensity plots displaying

the intensity values for a microarray before (left panel) and after (right panel) ANOVA

normalization.

1.2.2 Differential Expression Analysis (Feature Selection)

Often, the initial question of microarray data is which genes are differentially

expressed. Such DGEA is carried out by tests of inference including t-tests, ANOVA,

significance analysis of microarrays (SAM)[39], Mann-Whitney U test, and Wilcoxon's

matched pairs signed rank sum test[28,43,46,47]. Other than DGE, microarray data can be used

to study genetic regulation, discover gene function, and classify tissue samples.

For class comparisons, we are interested in whether the mean gene expression for

a given gene is significantly different between populations. The student's *t*-test is used for

two class comparisons. It can also be used to test against a constant such as two fold expression. The test statistic assuming equal variance in the two groups is,

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} - \frac{s_2^2}{n_1}}} \qquad [1.5]$$

where $t \sim t(df)$, $\bar{y}_i$ is the mean for all y i= 1-2 and s_i is the sample standard deviation for each i. Here we are testing the null hypothesis H_0: $\mu_1 = \mu_2$ *vs*. the alternative H_1: $\mu_1 \neq \mu_2$. For microarray data, the underlying distribution of gene expression values is often not normal, hence, the reference distribution is commonly estimated through permutation of the sample-wise gene expression values.

When there are more than two classes to be compared ANOVA can be used to test significance of DGE. The most commonly implemented ANOVA model in microarray analysis software is,

$$y_{ig} = \mu + \tau_{ig} + \varepsilon_{ig} \qquad [1.6]$$

which is a fixed effects gene-wise one-way model. The model assumes the $\varepsilon \sim N(0, \sigma^2)$ and are independent, τ and ε are also independent, where i = 1, 2,…,t, indicates the level of τ, and g indicates the gene. More complicated models such as [1.4] have been employed that seek to model the known sources of variation.

Here the model parameters are estimated using the method of least squares[58]. Briefly, the quantity to be minimized by least squares estimation is,

$$Q = \Sigma\Sigma(y_{ig} - \bar{\mu} - \tau_{ig})^2 \qquad [1.7]$$

The total variation associated with the data is given by the total sums of squares,

$$SSY = \Sigma\Sigma(y_{ig} - y_{..})^2 \qquad [1.8]$$

which can be further partitioned into sum of squares treatments (SST) [1.8] and sum of squares error (SSE) [1.9]

$$SSY = SST + SSE \qquad [1.9]$$

$$SST = \Sigma\Sigma(y_{i.} - y_{..})^2 \qquad [1.10]$$

$$SSE = \Sigma\Sigma(y_{ig} - y_{i.})^2 \qquad [1.11]$$

The test of significance is the F-statistic, which is given by

$$\hat{F} = (SST/df_T)/(SSE/df_E) \quad [1.12]$$

that is, the ratio of the mean squares treatment, divided by mean squares error. A mean squares quantity is simply the SS divided by its degrees of freedom. Here $\hat{F} \sim F_{i-1, E-1}$. With model [1.4] we are testing the null hypothesis: H_0: All $\tau_i = 0$ vs. H_A: some (at least two) of $\tau_i \neq 0$. Finally, the calculated F-statistic can be referenced to the F distribution or, as with the t-test, to a distribution estimated by random permutations of the data[28,38,43,58].

Multifactor ANOVA can be preformed as described in[28,43,54] utilizing the MAANOVA software by Wu, et al.. This R package provides an environment to construct fixed effects or mixed models. Tests of significance are carried out by computing *F*-ratios on model fitted residuals. Significance is assessed relative to an estimated distribution based on sample or residual shuffling54. For a reference design experiment, a model such as Eq. 1.13 would be appropriate to test for an effect due to the variety (i) of sample (i.e., testing the hypothesis that at least two of the sample groups will have significant differences in mean gene expression levels). This model, Eq. 1.13,

estimates significance by taking into account systematic sources of noise due to the array, dye, spot, and so on.

- Red indicates selected by the F2 test

- LOG 10 tabulated p value for the F1 test

Threshold for F3 test

Root mean square of the relative expression values

Significance threshold of the F1 test
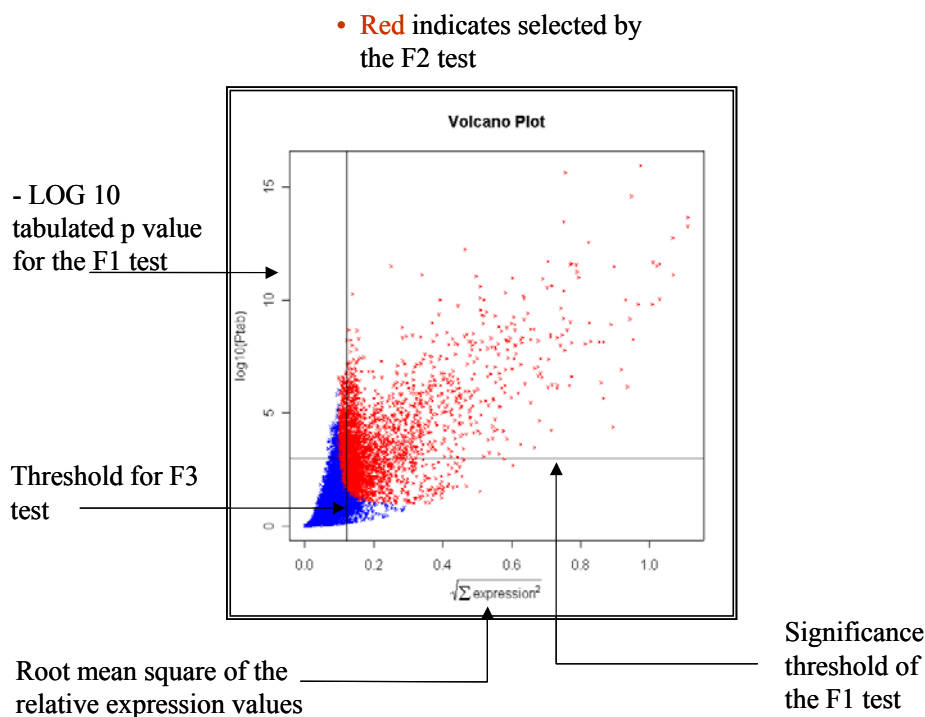
Figure 1.7. Volcano plot of the results of an F-Test. Significantly differentially expressed genes are located in the upper right hand corner. Graphical features are noted on the Figure.

$$H_o: y_{akghj} = \mu + A_a + D_k + G_g + AD_{ak} + VG_{ig} + DG_{kg} + AG_{ag} + S_{h(a)} + V_i + \varepsilon_{akghj} \qquad [1.13]$$

Here, μ is the mean of all spots on all arrays, A is the effect of array a, D is the effect of dye k, G is the effect of gene g, AD is the array times dye interaction, VG is the variety times gene interaction, DG is the dye times gene interaction, AG is the array times gene interaction, S is the effect of spot h, and V_i is the ith variety

(treatment/class/group). The genes declared as significantly differentially expressed via F-tests of model comparisons are displayed in volcano plots (Figure 1.7). Significance is assessed by computing three *F*-tests developed specifically for microarray data[54].

Perhaps the most versatile test of significance for microarray data is the so-called SAM[39]. This method can be employed for testing two classes, testing against a constant, testing for differences among multiple classes, or censored survival[45]. This method computes a value for $d(g)$, which in the case of dichotomous inquiries, is a relative distance, but can be replace by other functions such as the Cox proportional hazards function Eq. 3.9 for survival analysis or for changes in expression between three or more classes, *d(g)* can be defined in term of Fishers linear discriminant[39].

$$d(g) = \frac{\bar{x}_I(g) - \bar{x}_u(g)}{s(g) + s_0} \qquad [1.14]$$

$$s(g) = \sqrt{a\left\{\sum_m [x_m(g) - \bar{x}_I(g)]^2 + \sum_n [x_n(g) - \bar{x}_u(g)]^2\right\}} \qquad [1.15]$$

Where $\bar{x}_I(g)$ and $\bar{x}_u(g)$ are the average levels of gene (g) expression in states I and U respectively, s(g) is the standard deviation of the repeated expression measurements, Σ_m and Σ_n are summations of the expression measurements in states I and U respectively, *a* = (1/*m+n*)/(*m+n*-2), and *m* and *n* are the numbers of measurements in states I and U.

The t-test or pair-wise SAM were used for binary class comparisons, ANOVA and multifactor SAM were used for multiple class comparisons and SAM censored survival was used for survival analysis.

1.2.3 Class Prediction using Microarray Data

Many class prediction scenarios have been considered using microarray data[2,7,26,41,45,49,50,59-67] and essentially fall into the machine learning class of algorithms. A learning method is *supervised* if information regarding class labels or sample characteristics is supplied to the algorithm. Examples of these included k-nearest neighbors (*k*-NN), support vector machines (SVM), and Fisher's discriminant analysis. It is often necessary to remove genes from the data that have low situational relevance. In such cases, differential gene expression analysis such as ANOVA, weighted voting, or proportional hazards regression, can be used to select features (genes) that are most relavent[8,45,46,68], while other methods have been developed that effectively mine all of the data for strong predictors of relevant biological information[67]. Alternatively, unsupervised methods seek to find patterns in data without prior classification information. These methods include hierarchical clustering, terrain maps, principle components analysis (PCA), and the strong feature selection method by Kim et al.,[67].

Consider a gene expression data set containing *g* genes and *n* mRNA samples summarized as an *n x p* matrix *X,* where $x_{sg} \in X$ denotes the expression level of gene *g* in mRNA sample *s*. For each mRNA sample for which the class membership is known the data consists of the gene expression profile $x_s = (x_{s1},...,x_{sg})$ and a class label y_l where *l*= 1 to the number of tumor classes *k*. A class prediction algorithm ψ seeks to partition the space *X* in to *k* discrete subsets, $Z_1,...,Z_k$, such that for a sample with expression profile *x* = $(x_1,..., x_g) \in Z_k$, *k* is the predicted class[47,67,69]. We ultimately seek a classifier that is *consistent,* meaning that the expectation of the cost of estimation, $E(\Delta_s) \rightarrow 0$ as $s \rightarrow \infty$,

where $\Delta_s = \varepsilon_s - \varepsilon_{\bullet}$, that is, the error of the classifier ε_s of ψ_s minus the Bayes error ($\varepsilon_{\bullet}$) for s samples[70]. In other words, we would like the classifier to be achieve low error for the entire real population given the limited training data.

When sufficient data is available, prediction parameters may be specified from a training set (T) that contains known class distinctions such that $T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_s, y_s)\}$. The prediction parameters (i.e.: weight, distance, similarity) can then be applied to a unknown set L such that $L = \{\mathbf{x}_1, \ldots, \mathbf{x}_s\}$ to predict the class (y_k) of the observation x_g in L set[47].

When y_k is known, the prediction and true classes can be compared to estimate the error rate of the predictor[47] by methods such as leave-one-out cross validation (LOOCV) or adding random error[67]. Here, the classifier is run on the data, but with one sample left out, the classification rule is calculated for the remaining samples and is used to predict the class of the omitted sample. Each sample $\mathbf{x}_s$ is, in turn, left out and cross-validated. The resulting classification for each member in the training set is then compared to the true classification and an error rate is determined. This method works well for large sample sizes and is unbiased, but can be over optimistic when samples are limited due to increased error variance[67].

Algorithms such as the signal-to-noise ratio (S2N)[2,7], the independently consistent expression discriminator (ICED)[49], k-NN[2,47], SVM[49], gene shaving (GSH)[50], and strong feature set determination (SFSD)[67,71] have been used successfully in class prediction via gene expression. In the context of brain tumor research, *k*-NN and SFSD have been

applied to find a limited panel of genes capable of reclassifying histologically ambiguous tumors or distinguishing binary grade comparisons respectively[7,67].

One of the first published algorithms specifically developed for prediction using gene expression data was the S2N[2,7]. The ICED, inspired by the S2N, was demonstrated to be more accurate in finding a dichotomous classification rule compared to analysis by SVM and k-NN[4]. Both the S2N and ICED, however, have the disadvantage of being binary classifiers. The other predictors introduced above (k-NN, SVM, GSH, SFSD), can be used to address classification problems where there are greater than two classes although *our* current implementation of SVM and GSH are for binary classifications. Perhaps the most rigorous reported method was SFSD, an algorithm that finds feature sets by first increasing the variance of the expression measures then uses a heuristic guided random walk search algorithm to identify gene sets (three at most) that achieve a low error rate[67]. Because microarray data contains large numbers of features combined with multiple samples and sample classes, finding all optimal feature sets is computationally intractable. Thus, researchers must rely on sub-optimal feature sets, which may nevertheless perform very well by achieving an appreciably low error rate while keeping in mind Occam's razor; it is preferable to utilize a simple function that explains most of the data than a complex one[69].

A caveat to class prediction based on histological stage is that classification of the samples used to train the model is subject the same histological classification error present during the original classification. Samples that are, as such, incorrectly classified are used to construct the classification parameters of the predictor, leading to bias. A way

to circumvent this pitfall is to reclassify the samples into some other meaningful way, such as by patient survival[8] or in the case of GSH, set constraints that include class relevant information[50]. Despite the challenges, gene expression profiles undoubtedly contain information that can be used to construct prognostic estimates, and eventually may supplant histological classification as the standard for diagnostics.

The following sections detail some common analysis and prediction techniques employed in the microarray field. In the case of S2N, gene shaving, SVM, and SFSD, feature sets are selected explicitly during the first stages of computation. For classifiers such as *k*-NN, feature sets are usually chosen by some 'outside' algorithm such as ANOVA or proportional hazards regression.

1.2.3.1 Signal-to-Noise Ratio

One of the earliest and most cited reports attempting to identify genes most that are good predictors of class membership is the weighted voting algorithm developed by Golub et al., known subsequently as the S2N[7]. This algorithm was initially shown to

$$P_g = [\mu_1(x_g) - \mu_2(x_g)]/[\sigma_1(x_g) + \sigma_2(x_g)] \quad [1.16]$$

predict class membership between two types of leukemia. Their measure of "correlation" Eq. 1.16 is a minor variant of a special case of *sample maximum likelihood* discrimination rule

$$C(x) = \arg\min_l \sum_{s=1-n} \frac{(x_s - \mu_{ls})^2}{\sigma_s^2} \qquad [1.17]$$

which has also been referred to as the diagonal linear discriminant analysis rule[47]

Eq. 1.17. This rule is evoked in when the class densities have the same diagonal

covariance matrix $\Delta = \mathrm{diag}(\sigma^2_1,...,\sigma^2_n)$. For classes $k = 2$, the sample maximum likelihood

rule classifies an observation $\mathbf{x} = (x_1,\dots,x_l)$ as 1 iff

$$\sum_{g=1-l} \frac{(\bar{x}_{1g} - \bar{x}_{2g})^2}{\hat{\sigma}_g^2} \left(x_g - \frac{(\bar{x}_{1g} + \bar{x}_{2g})}{2} \right) \geq 0 \qquad [1.18]$$

which can be rewritten as $\Sigma_g v_g$, where $v_g = a_g(x_g - b_g)$, $a_g = (\bar{x}_{1g} - \bar{x}_{2g})^2 / \hat{\sigma}_g^2$. Here a_g is

almost the same function used by Golub, et al., Eq. 1.16 except that the denominator is

the sum of the standard deviation of gene g instead of the variance[47].

The correlation metric $P_g = a_g$ is used to weight the vote (v) function Eq. 1.18

which is the weighted vote for gene g. Further, $\Sigma|v_{+j}| = V_{\text{class 1}}$, and $\Sigma|v_{-j}| = V_{\text{class 2}}$, such that

the prediction strength for a sample under test is given as PS = V_{class1}- V_{class2} / (V_{class1}+

V_{class2}). Significance was estimated by comparing the predicted classes to predictions

made by random permutations of the sample labels.

The S2N algorithm was initially applied to a leukemia data set that consisted of

72 total leukemia samples, some ALL and others AML. After training the algorithm with

38 samples, class membership was predicted among the remaining 34 samples. The

authors reported greater than 85% accuracy in predictions with various numbers of the

highest weighted genes[7].

1.2.3.2 ICED Analysis

Bijlani, et al., have recently (2003) developed an algorithm for distinguishing two classes of samples and a finding a minimum number of predictor genes[49]. The algorithm: Independently Consistent Expression Discriminator (ICED), is loosely based on a measure of correlation (Pearson's) followed by calculation and ranking of weighted votes. It is inspired by the S2N ratio of Golub, et al., but was demonstrated to perform better in terms of accuracy of prediction for the leukemia data set. ICED also allows for the possibility of variable gene expression for a gene in one class but constant expression of the same gene in the other class, while the S2N ratio would not likely vote such as gene as a good discriminator. The equations for the algorithm are listed below Eqs. 1.19-1.23. Here the subscripts m,n denote the number of samples in each class, g denotes gene (g), and g^* indicates gene (g) in unknown sample g*.

ICED can be summarized in four steps to train the algorithm and three steps to validate the predictors. The four training steps are: *i)* format data, *ii)* normalize (mean centering), *iii)* rank each gene by weight statistic Eqs. 1.19, 1.20,

$$W_{1g} = \frac{\frac{1}{m} \sum_{i=1-m} | x_{2,g} - \mu_{1,m}(g) |}{\sigma_{1,n}(g)} \quad [1.19]$$

$$W_{2g} = \frac{\frac{1}{n} \sum_{g=1-n} | x_{1g} - \mu_{2,m}(g) |}{\sigma_{2,m}(g)} \quad [1.20]$$

to indicate its usefulness as a predictor, where m is the number of samples in class 1, n is the number of samples in class 2, for gene g, *iv)* find optimal voters. Prediction consists

of *i)* format unknown sample data, *ii*) classify unknown samples using calculated votes Eqs. 1.21, 1.22,

$$V_{1g} = W_{2g} \bullet \left| g^* - \mu_{2TR,m}(g) \right| \quad [1.21]$$

$$V_{2g} = W_{1g} \bullet \left| g^* - \mu_{1TR,n}(g) \right| \quad [1.22]$$

and *iii*) assign class membership. The strength of prediction is given by Eq 1.23,

$$P(g) = \frac{q \bullet \sum\limits_{g=1-p} V_1(g_g) - p \bullet \sum\limits_{g=1-q} V_2(g_{g_i})}{q \bullet \sum\limits_{g=1-p} V_1(g_g) + p \bullet \sum\limits_{g=1-q} V_2(g_{g_i})} \quad [1.23]$$

which falls in the interval [-1,1]. Large values of *P(g)* indicate greater prediction strength.

It is easily seen from inspection of Eqs. 1.21, 1.22 that the weight of a gene in class-one is generated with respect to the standard deviation of the same gene in class-two. The authors suggest that the voting equations are the same as those in Golub, et al., however, the equation $v_g = a_g(x_g - b_g)$ is derived from the *sample maximum likelihood discriminator*[47] where $b_g = (\bar{x}_{mg} - \bar{x}_{ng})/2$. In Eqs. 1.21,1.22, $b_g = \mu_{iTR,m(n)}$, where $\mu_{iTR,m(n)}$ is simply the mean of gene (*m* or *n*) in class *i*. Nevertheless, the authors demonstrate remarkably accurate predictions using the Golub, et al., data set, as well as a much smaller data set based on a mouse model of Batten disease[49].

This method was reported to have several advantages over the S2N statistic for binary classification. First, it can be used on a small number of samples. The authors were able to accurately predict biologically significant genes from a mouse model of Barrett's disease with only eight total samples. Secondly, it was able to more accurately

predict tumor classification for the Golub et.al.data set than the S2N statistic, Support Vector Machines, and Neighborhood analysis. One drawback to this method is that it can only be used to distinguish two classes, and multi-class problems are out of the range of this approach.

1.2.3.3 Nearest Neighbor Classifiers

Nearest neighbor classifiers are a simple and powerful class of algorithms that are based on a similarity (or distance) function between observations; in this case, class-specific gene-wise average intensities. One of the most popular among these is k-NN. For expression profiles x = $(x_1,…,x_g)$ and x' = $(x'_1,…,x'_g)$ the degree of correlation, for example, is based on a correlation coefficient[47] such as given in Eq. 1.16. Where r is the correlation measure, x_g is the expression level of the gth gene in class one, x'_g is the expression level of the gth gene in class two, and $\bar{x}$ ($\bar{x'}$) is the mean expression level of class one (two). The k nearest neighbor rule is computed as follows, (i) find the k closest observations in the training set, and (ii) predict the class that is most common among k neighbors[47,59]. The number of k neighbors can be specified by leave one out cross validation (LOOCV) by performing for a number of k's and retaining the one (k) with the smallest error rate. The number of classes in T is specified *a- priori*. Finally k-NN classifier is universally consistent[70] if $k \to \infty$ in such a way that k/n $\to$ 0 as n $\to \infty$.

This method was used by Nutt et al., predict membership of histologically ambiguous tumor specimens as either oligodendroglioma, a somewhat benign cancer with favorable prognosis, or astrocytoma, a lethal brain cancer with poor prognosis. They demonstrated a maximum accuracy of about 86% after constructing the classifier with 20

genes. It was concluded that the gene expression data predicted prognosis more accurately than histopathological classification.

1.2.3.4 Support Vector Machines

One class of *modern* techniques for data analysis are machine learning tools, such as support vector machines (SVM), that seek to find a linear discrimination rule for data with high dimension by non-linear re-mapping of the data into higher dimensional space.

$$\Phi: X^N \to F$$

$$x \mapsto \mathbf{x} := \Phi(x)$$

The algorithm seeks a function that can partition F into a dichotomous space:

$$F \to y_i \in \{\pm 1\}$$

X^N is a input data space with dimension N and F is the feature space. In F a simple (linear) discriminant rule (hyperplane) can be applied that would not have been
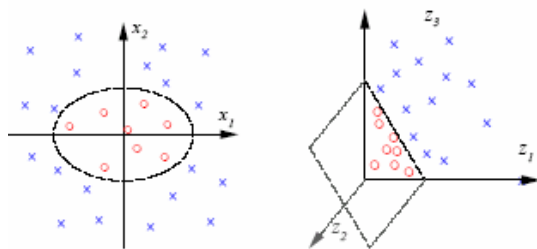
Figure 1.8. Depiction of a 2D data space re-mapped to a 3D space. In the left example, the data are arranged in a 2D space and one class (red) is surrounded by the other (blue). On the right a hyperplane can be specified by a linear function to separate the classes.

successful in X^N data space[69]. Consider the example in Figure 1.8. On the left is a 2D space that contains the data, clearly a linear discrimination rule would not be able to separate the classes in the data. However, if the data are re-mapped to a 3D space, then a linear discrimination rule can be applied to separate the two classes.

One challenge with this approach is that the dimensionality of F can increase drastically as N increases, making computations intractable even for simple discriminant functions. For certain feature spaces however, *kernel functions* can be used to compute scalar products between data points. Some common kernel functions are listed in Table 1.1. In the binary classification setting the decision rule is given by

$$\phi(\mathbf{x}) = \text{sign}[f(\mathbf{x})]$$

where for data and class labels (x_i, y_i), i= 1,…,n a function, $f(\mathbf{x}) = h(\mathbf{x}) + b$, is sought with $h \in \mathbf{H_K}$ (a reproducing kernel Hilbert space (RKHS)) and b a constant minimizing Eq.1.24

$$\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \parallel h \parallel_{H_K}^2 \quad [1.24]$$

where $(x)_+ = \max(x, 0)$, $\parallel h \parallel^2_{Hk}$ is the square norm of the function h defined in the RKHS

Table 1.1: Some examples of kernel functions taken from Muller, et al.

Gaussian RBF	$k(\mathbf{x}, \mathbf{y}) = \exp\left(\dfrac{-\parallel \mathbf{x} - \mathbf{y} \parallel^2}{c}\right)$
Polynomial	$((\mathbf{x} \cdot \mathbf{y}) + \theta)^d$
Sigmoidal	$\tanh(\kappa(\mathbf{x} \cdot \mathbf{y}) + \theta)$
inv. multiquadric	$\dfrac{1}{\sqrt{\parallel \mathbf{x} - \mathbf{y} \parallel^2 + c^2}}$

with the reproducing kernel function $K(\bullet,\bullet)$ which measures the complexity or smoothness of h. Finally, λ is a tuning parameter which balances the data fit and complexity of $f(\mathbf{x})$[66].

SVM is traditionally a binary classifier but it has been modified for prediction of multiple classes (multi-class support vector machines (MSVM)) and used successfully in class prediction problems with gene expression data[66] using the data set of Golub, et al.. Lee, et al., achieved between 11% and 3% training error depending on the number of genes used, the choice of kernel function and the preprocessing steps for the input data. The authors also examined a data set with four classes of cancer. MSVMs correctly classified 100% of the test samples.

Thus, SVMs and MSVMs are a viable choice for tackling the problem of using gene expression data for class prediction. Our current implementation of SVMs is as a binary classifier only.

1.2.3.5 Gene Shaving

A relatively simple and intuitive method of class prediction using gene expression data is so called gene shaving. The basic concept of this algorithm is to find k-subsets S_k of genes that maximize the variance of the gene average Eq. 1.25,

$$\text{var}(\frac{1}{k}\sum_{i \in S_k} x_g) \quad [1.25]$$

$$D_k = \frac{V_{Between}}{V_{Total}} = \frac{\frac{1}{l}\sum_{g=1}^{l}(\overline{x}_g - \overline{x})^2}{\frac{1}{lk}\sum_{i \in S}\sum_{g=1}^{l}(x_i - \overline{x})^2} \quad [1.26]$$

as well as genes that show high coherence[50].

As described in Hastie, et al., the algorithm consists of seven steps: *i*) beginning

Figure 1.9. Gene shaving cluster formation
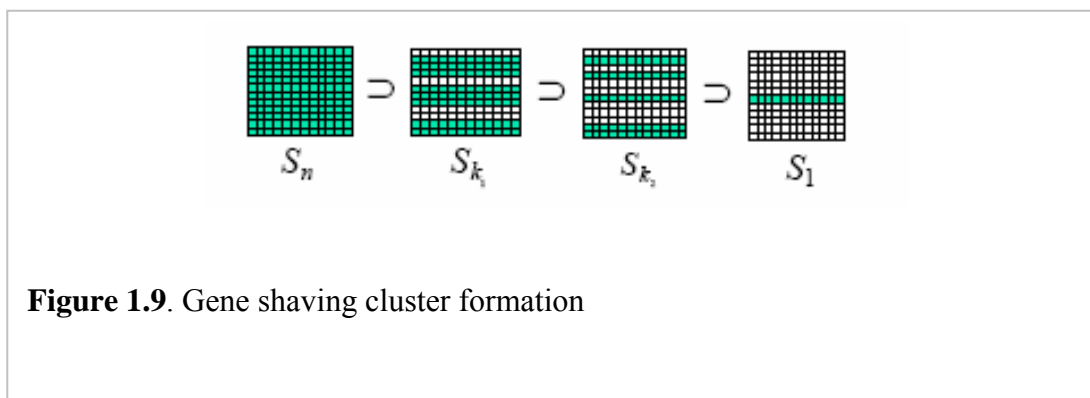
with the entire expression matrix *X,* each row is centered to have a zero mean *ii*) compute

the leading principle component of the rows of *X iii)* shave off the proportion α

(typically 10%) of the gene having the smallest absolute inner-product with the leading

principle component *iv*) repeat steps *ii* and *iii* until only one gene remains, *v*) this

produces a nested sequence of gene clusters $S_N \supset S_k \supset S_{k1} \supset S_{k2} \supset,\ldots,\supset S_1$ where S_k

denotes a cluster of k genes for which the optimal cluster size is estimated using the gap

statistic [18], *vi*) orthogonalize each row of *X* with respect to $\overline{x_{sk}}$, the average gene in $\hat{S}_k$

,and *vii*) repeat steps 1-5 with the orthogonalized data to find the second optimal cluster.

This process is continued until a maximum of M clusters are found where M is chosen *a*

priori[50].

The first paper to introduced this learning method was that of Hastie, et al.. In the manuscript, the authors describe the algorithm and assert that gene shaving can be preformed in an unsupervised, supervised or partially supervised manner. They go on to report the supervised approach for predicting survival in patients with large B-cell lymphoma.

1.2.3.6 Selection of Strong Feature Sets

The goal of extracting genes that are strong predictors of a relevant biological class is the object of many classification algorithms. However, for microarray experiments, sample sets are often small and comprise only a few members of each relevant class. Kim, et al., employed the supercomputer facilities at the NIH and developed an algorithm based on a perceptron that finds strong feature sets for class prediction. Classifiers were designed from a probability distribution created from spreading the distribution of the expression measures in a circular fashion to increase the difficulty of classification. The algorithm was parameterized by the variance of the circular distribution and the goal is to find gene sets whose classification accuracy remains strong despite increased spreading of the sample data. In this case the error, as a function of the inflated variance, gives an indication of the strength of the feature set[67,70,71]. An example is given in Figure 1.10, taken from[67].

(a) no spreading   (b) spreading with $\sigma = 0.3$

(c) spreading with $\sigma = 0.6$ (d) spreading with $\sigma = 1.0$

Figure 1.10. Illustration of the sample spreading method for identification of strong feature sets. From A to D, sample variance is increased and a classification rule is computed. Note that if spreading causes the error to increase substantially, than the feature set is not regarded as strong.

Feature sets are identified by a heuristic search algorithm that proceeds through a guided random walk, and is of a class of algorithms known as preceptrons. If a feature is a member of an acceptable solution set containing a small number of features then it is more likely to be a part of an acceptable solution using a larger set. Such algorthims can be considered genetic search or stochastic[67,70,71]. In mathematical terms, identification of

the optimum set of genes to predict class membership would require an essentially infinite number of comparisons when *n* and *g* are both large. This method of finding strong feature sets is useful for finding many good solutions rather than finding a best solution[70].

1.2.3.7 Prediction Analysis For Microarrays

This method, developed by Tibshirani, et al., performs automatic feature selection by shrinking the class specific centroids[72]. Classification is similar to *k*-NN except that class membership is determined by distance to the class specific centroid. Cross-validation is used for generalization error estimation. Prediction Analysis of Microarrays (PAM) will be discussed in greater detail in Chapter 3.

1.3 GLIOMA BIOLOGY AND GENETICS

The central nervous system (CNS) is comprised of two classes of cells: neurons and neuroglia (glial cells). Neurons can be further categorized by function into motor, sensory, and interneurons[73], and are the information processing cells of the nervous system. Glial cells outnumber neurons in the central nervous system 10 to 50 times[74] but they do not conduct nerve impulses or have a direct information processing role, rather they play a support role for neurons[73]. Glial cells can be divided into macroglia (astrocytes, oligodendrocytes, and ependymal cells) and microglia which are sometimes phagocytic[74].

Generally star shaped, astrocytes (astro= star) (cytes = cells) are the most numerous of glial cells filling almost all of the extraneuronal space[73,75]. Astrocytes have many functions including regulation and storage of potassium (K^{++}) and removal of

neurotransmitters such as γ–aminobutyric acid and serotonin from the local environment[74,75]. Astrocytes can be subdivided as protoplasmic or fibrous depending on the presence or absence of cytoplasmic fibers[75]. Protoplasmic astrocytes are predominately found in the gray matter while fibrous astrocytes are primarily found in the white matter[75]. Oligodendrocytes produce the myelin sheaths that surround axons in the CNS. A single oligodendrocyte can myelinate between 10-15 axons doing so by spiraling around the axon during neural development[75].

Astrocytomas (ASTs) and oligodendrogliomas (ODGs) are neoplasms that stem from astrocytes and oligodendrocytes respectively. For World Health Organization (WHO) tumor grading occurs on a malignancy scale from one to four. In the case of glioblastoma multiforme (GBM), (WHO grade IV), specimens have been collected that have features of both AST and ODG tumors. In terms of prognosis, the ODGs is more favorable than ASTs[2,76], so accurate classification of histologically ambiguous tumor specimens is of paramount importance. Gliomas generally cause symptoms (e.g. seizures) by perturbing cerebral function, elevating intracranial pressure by either mass effect or obstructing cerebrospinal fluid (i.e. hydrocephalus), or causing neurologic (and sometimes endocrine) abnormalities (e.g. paralysis, sensory deficits, aberrant behavior, headaches)[77-79].

A hallmark of AST tumors is resistance to apoptosis, and by extension to most current chemotherapeutics and radiation[14] and rapid progression. If patients with resistant tumors are given standard therapies, they suffer ineffective treatment, lower quality of life, and sometimes devastating economic losses. There is ample evidence that surgery

confers little if any survival benefit and adjuvant therapies have been similarly

unsuccessful[14] .Microarray technologies provide a powerful way to understand two

crucial pieces of information necessary to improve treatment of this disease: objective

classification, and the ability to obtain provide a genetic signature that can be correlated

with treatment response.

Oligodendrogliomas are associated with WHO malignancy grades II and III, the

latter being anaplastic oligodendroglioma (AOD). This tumor mainly occur in adults and

OGDs are relatively benign, however, progression to AOD can occur which carries a less

favorable prognosis. Classic ODG tumors are characterized by moderate cellularity, little

mitosis, no necrosis, and have a 'chicken wire' capillary morphology[76]. (Appendix A1).

They are usually not invasive and recurrence occurs at the primary site. The more

advanced AOD is characterized by increased nuclear pleomorphism, hyperchromatism,

hypercellularity, prominent microvascular proliferation and necrosis. Genetic lesions

include gene deletions from chromosomes 1p and 19p, over expression of *EGFR*, *PDGF,*

and *PDGFR*[76].

Astrocytoma tumorigenesis has been proposed occur via two genetic pathways *i)*

a *de novo* pathway in which a high-grade astrocytoma develops without a previous tumor,

and *ii)* a progression pathway during which a high-grade tumor develops from a low-

grade precursor (II, or III). Glioblastomas can also arise *de novo* or from a progression

from a lower grade[80]. Some GBMs show predominately astrocytic features while others

show more mixed AST and OGD features[76], which begs the philosophical question of

cellular origin and / or, whether certain cells in the neoplasm re-differentiate to produce a different glial phenotype?

Primary ASTs can be characterized by a high frequency of *EGFR*-gene amplification and a low frequency of *p53*-gene mutation. *EGFR* gene amplification occurs most frequently in glioblastomas associated with loss of a complete copy of chromosome 10[13,80]. Indeed, WHO subdivisions of high-grade astrocytomas (III and IV) have been made on the basis of frequently found genetic changes such as, *p53*-gene mutation, loss of heterozygosity on chromosome arm 17p (LOH 17p), LOH 10 and EGFR-gene amplification. Secondary / progressive tumors show a high incidence of p53 mutation, a low incidence of *EGFR* amplification and eventually LOH 10[13]. Further, *p53* mutations have been identified in 60–80% (or more) of low-grade astrocytomas. This mutation appears to have higher incidence in young patients (ages 18-40)[13] *IGFBP-2* has consistently been found to be over-expressed in GBMs, and six genes, including *TIMP3*, *EGFR*, and *GDNPF*, have been found to be over-expressed in 64–100% of grade II tumors[41]. Seven genes, including *PDGFR-α, PTN*, *LRP*, and *SPARC*, were up-regulated by at least 2-fold in 20–60% of grade II tumors[41]. Elevated expression of the *EGFR*, *MDM2, CDK4, CD44, IGFBP2, DAP-3*, and *laminins* is well described by microarray studies of gliomas [2,4,15,41,48,81]. Leenstra, et al., utilized molecular techniques to sub-type AST tumors (63 GBM 12 AA), for loss of heterozygosity on chromosome 10 and p53, and *EGFR* amplification. They reported the results of Cox proportional hazards modeling revealed that age and genetic subtype were significant prognostic indicators while histological grade was not.

Mariani, et at., demonstrate that *DAP-3* was induced in the invasive rim of Glioblastomas, and that there is considerable heterogeneity of gene expression across the GBM tumor mass[4].

The evolution of this reductionism is in utilization of high throughput technologies such as the DNA microarray where the application of machine learning can take advantage of the massive amounts of data. In contemporary terms, DNA (and protein) based molecular profiling devices are poised to significantly impact the way medicine is developed and administered.

1.4 CURRENT GLIOMA CLASSIFICATION METHODS

The Kernohan, St. Anne/Mayo (SAM-A), World Health Organization (WHO) and TESTAST 268 protocols are the most commonly used 4-tier grading systems for classifying grade and stage of astrocytomas, none of which is universally accepted[82]. This situation has obvious inadequacies, hindering prognostic assessment, comparative evaluation of tumors, and inter-center data comparison, while contributing to generalization of therapy, subjective diagnosis, misdiagnosis, unnecessary medical costs, and procedures. Survival curves generated by Karak, et al., for each of these grading protocols were similar suggesting that results obtained by any one of the protocols can be generalized to the others[82]. Interestingly, intra-classification grade-wise survival analysis revealed differences between grades 2 and 3 or 4 but not between 3 and 4. Despite this reported correlation between classification methods themselves, estimates of error rates as high as 30% have been demonstrated in the literature[82], decrying a need for improved classification methods based on parameters that can be objectively defined.

The World Health Organization (WHO) scheme is based on the appearance of certain characteristics: atypia, mitoses, endothelial proliferation, and necrosis. These features reflect the malignant potential of the tumor in terms of invasion and growth rate. Tumors without any of these features are grade I (pilocytic astrocytoma), and those with one of these features (usually atypia) are grade II (low grade astrocytoma). Tumors with 2 criteria and tumors with 3 or 4 criteria are WHO grades III (anaplastic astrocytoma) and IV (GBM), respectively. Thus, grades I and II are the low-grade group of astrocytomas[78,79,83,84,85]. Example images of the most common gliomas are shown in appendix 1 (Table A1).

Glioblastomas are known to occur predominantly throughout the cerebrum with infiltrative processes that can extend to the contralateral hemisphere[14]. Infiltrating low-grade astrocytomas tend to occur in the lobes of the cerebral hemispheres, especially in the frontal lobe. Pilocytic astrocytomas may occur in the frontal, temporal, and parietal lobes and cerebellum, but they are also common in locations closer to the midline, such as the hypothalamus, thalamus, optic chiasm, and brain stem[77,79,85]. In children, pilocytic astrocytomas have a tendency to occur in the mesial structures of the cerebellum[14,77,79,85].

Due to their remarkable pathology, a subset of astrocytomas comprised of juvenile pilocytic astrocytoma (JPA), pleomorphic xanthoastrocytoma (PXA), and subependymal giant-cell astrocytoma (SGCA), are not effectively classified by a 4-tiered grading system such as the WHO. These tumors can have endothelial proliferation as well as cellular atypia. Fortunately, they are slow growing and well-defined making surgery curative in most cases[14,77,79,85].

1.5 DNA MICROARRAYS AND BRAIN TUMOR RESEARCH

Initial microarray studies of gene expression in gliomas identified differentially expressed genes and established gene panels that were distinctive relative to histopathological class[2,23,67]. Later studies focused on finding genes that could predict tumor class[2,23,67]. Two reports have related gene expression patterns with survival in gliomas[2,23]. Genetic correlates with survival have been described in other genetic (non-microarray) studies involving gliomas[81]. In all, eleven reports have been published detailing brain tumor genetics through microarray analysis. These findings validate the hypothesis that gene expression can be used to identify new tumor subclasses, yield novel therapeutic targets and provide highly accurate diagnostic and prognostic information[2,6,13,14, 23,25,77,79,81,85,86]. An important area left for consideration is how gene expression relates to survival and other outcomes. Nutt et al., addressed this question from the important standpoint of classifying ambiguous tumors into more appropriate histological categories that were more accurate predictors of patient survival, but they did not relate gene expression patterns directly to length of survival[2]. The question at hand is, given a particular gene expression pattern, how long is the patient likely to survive, and which genes can most reliably answer this question. This information will promote three advances, *i*) design of DNA based diagnostic devices, *ii*) indication of novel pharmaceutical targets for improved therapy, and *iii*) enhancement of the ability of the clinician to plan and manage personalized treatment.

1.5.1 Basic gene expression analysis

Three studies involving glioma genetics made histological comparisons and found differential expressed genes simply by calculating mean expression ratios and reporting those that were beyond a given threshold. One of the first microarray studies of glioma genetics was conducted by Ljubimova, et al., They studied a total of 12 tissue samples that included 5 GBMs, 2 AAII, 1 meningioma, and 2 normal brain tissues. One of the key findings was the identification of expression patterns of *laminin*-8 and *laminin*-9 that could be correlated with time to tumor recurrence for GBMs. In addition they detected 2345 genes with increased expression and 719 genes with decreased expression compared to normal brain. Of these 14 were up regulated > 2-fold in all 5 GBMs. They further demonstrated that tissue adjacent to GBM had only slight differences compared to normal brain but that *Laminin α 4 chain, keratin18,* and *Desmoplakin* were all up regulated compared to the GBM tumors.

Sallinen, et al., utilized microarrays and tissue chips to identify differentially expressed genes in seven astrocytomas, three GBMs, and two specimens that represented a primary and recurrent grade III astrocytoma[20]. The microarrays, from Clontech consisted of 588 genes and were hybridized with [α-^{33}P]dCTP labeled RT product. The authors also prepared a tissue microarray composed of 418 individual tumor samples (364 gliomas and 54 other types of brain tumor). They reported prominent changes in expression fold change but did not perform any statistical tests of significance. Notable genes reported induced in GBM vs. normal brain included *SPARC*, *Timp-1*, *Timp-2,* c-myc, *vimentin*, *VEGF,* and *TGF-β.* In addition the authors reported 10 genes

differentially expressed between primary *vs.* the recurrent sample. Tissue chip analysis revealed that *ISGFBP-2* status was significantly inversely correlated to patient survival[20]. *ISGFBP-2* and *vimentin* expression characteristics as reported by the microarray analysis were corroborated by the tissue chip immunohistochemical staining data indicating that comparison of large average fold changes was sufficient to reveal true differences in gene expression.

Markert, et al., demonstrated patterns of differential gene expression between four GBMs and three normal brain specimens[15]. Analysis was conducted by determining genes that were induced or repressed above a fivefold expression threshold. Using an Affymetrix GeneChip, they identified 34 of ~7,000 transcripts that were fivefold induced in all GBMs relative to normal brain. These genes included *p53-associated protein*, *MDM2, ISGFBP-5,* and *ISGFBP-6*. They also investigated the functional manifestations of the gene found to be differentially expressed by whole cell patch clamp. The microarray data generated the hypotheses that voltage-gated K^+ channel β3 subunit and NMDA receptor-activated currents would be down regulated in GBMs compared to normal brains. The authors demonstrated that the electrophysiological characteristics were consistent with the microarray findings validating the conclusions of the microarray analysis.

These studies demonstrate that simple fold-change analysis is an effective tool for identifying differentially expressed genes. However, they have the disadvantage that there is no way, outside of testing each observation, of determining the number of false positives, similarly there are undoubtedly many genes that are regulated below the fold-

change threshold that may be nonetheless significant. Further, this type of analysis does not take into account that patterns of gene expression may identify important regulatory networks, pathways, and molecular signatures.

1.5.2 Histological classification using microarray data

Histopathology has a long and successful history identifying many disease subtypes and relating them to therapeutic strategies and clinical outcomes. Thus, most microarray reports dealing with gliomas based their analysis on identification of differences between defined histological classes. The methods for this type of analysis include statistical inference, and supervised and unsupervised learning. A few groups have developed custom algorithms to circumvent challenges posed by the inadequacies of contemporary analytical techniques.

Huang et al., used cDNA Clonetech microarrays from identify differentially expressed genes in 11 low grade astrocytomas relative to normal tissue[6]. A students unpaired t-test was used to assess significant differential expression. Of the 1176 probes represented on the array they found 24 genes to be differentially expressed relative to normal tissue. These genes included *tissue inhibitor of metalloproteinase* (*TIMP3*), *epidermal growth factor receptor* (*EGFR*), c-myc oncogene, *Glia derived neurite promoting factor* (*GDNPF*), nm23-H4, AAD14, *60S ribosomal protein LS* (*rp*LS), *Low density lipoprotein receptor related protein* (*LRP*), SPARC, *h*BAP, *pleitrophin precursor* (*PTN*), PDGFR-α, *interferon-inducible protein 9-17* (*IFI 9-27*), *protein kinase CLK*, *teratocarcinoma-derived growth factor* (*TDGF1*), *GRB associated binder-1* (*GAB1*), *box-dependent myc-interacting protein 1* (*BIN1*), *Tyrosine protein kinase SKY* (*TYRO3*) ,

lactate –dehydrogenase-A (*LDH-A*), *Adducin 3* , guanylate kinase (*Guk1*), *keratin type II cytoskeletal 8* (*KRT8*), and *CDC10 protein homologue* (*CDC10*)[6].

Rickman, et al., found 360 genes to be differentially expressed between grade IV and grade I tumors by at least 1.5-fold in mean intensity ($P < 0.01$), 167 had increased and 193 had decreased expression levels in grade IV tumors *vs*. to grade I tumors, 183 genes were expressed at a higher level in grade IV relative to 5 grade II astrocytomas, and 703 genes were over expressed in glioblastomas compared with normal brain. Five genes (*ZYX, SDC1, FLN1, FOXM1*, and *FOXGB1*) were characterized that had not been previously associated with glioblastoma[41]. Significant differences between the mean normalized intensities was determined by one-way ANOVA Hierarchical clustering as preformed to visualize the differences in expression as a function of tumor histopathology.

Kim, et al., developed an algorithm for selecting histological class predictor genes in groups containing one to three members. The aim was to use a small sample set (25 tumors (10 GBM, 4 AA, 5 AO, 6 OGD)) and achieve superior classification error rate. This challenge was executed by a novel process of spreading the variance of the expression measurements for a given set of genes (3 at most)[67]. The algorithm they employed is detailed in section 1.2.3.6. They demonstrated classification rules for separating one class, say OGD, form the remaining classes, thus 4 classifiers were developed for each set of predictor genes. While not all gene panels were reported, it is interesting to note that there were fewer GBM discriminating genes panels, perhaps reflecting the large degree gene expression variability in this group of

astrocytomas[23,42,48].

Van den bloom, et al., tested for genes relevant to tumor progress in 8 samples of primary (grade II) vs. recurrent astrocytomas (grade III or IV)[86]. Sixty six genes were reported significantly different for P < 0.01, and ≥ 2-fold change in expression. A total of nine of these were corroborated by further analysis which included *COL4A2*, *FOXM1*, *MGP*, *TOP2A CENPF*, *IGFBP4*, *VEGFA*, *ADD3*, and *CAMK2G*. It was suggested that these gene play a role in tumor progression. Statistical inference was preformed by paired *t*-tests between sample. The population distribution was estimated by permutations. Interestingly, they reported RT-PCR fold change data and microarray fold change data for 15 genes. Fold change measurement form these technologies corroborated well in magnitude, contrary to reports suggesting DNA microarrays underestimate fold change [87,88,89].

1.5.3 Survival Classification using Microarray Data

Prediction of patient survival by gene expression profiling represents a powerful and important use of microarray technology. Currently, three glioma microarray studies have reported gene panels that related in some way to survival. Methods used for this type of analysis include, S2N, unsupervised clustering and class prediction. Shai et. al reported the identification of molecular subtypes of gliomas by analysis with Affymetrix GeneChips. They surveyed 35 glioma samples including ASTs, GBMs and OGDs. The authors used the Affymetrix U95Av2 chip and conducted multiple analyses using the S2N, t-test, multidimensional scaling, k-means and hierarchical clustering analysis and were able to find genes that partitioned the samples into all relevant comparisons

(primary *vs*. recurrent; astrocytoma *vs*. oligodendroglioma, 1yr survival *vs*. 3 yrs survival, and so on)[23]. For survival analysis, the authors preformed a *t*-test on samples that survived 1 year or less vs. survivors of greater than three years and coupled this data with the S2N algorithm to find predictor genes. Error was estimated by cross validation, and an error rate of 22% was reported for the survival comparison.

Nutt, et al., demonstrated that tumors with ambiguous histological features could be accurately re-classified into a histological class for improved prognostic accuracy (survival). Gene candidates for prediction modeling were determined by the S2N algorithm of Golub, et al.. k-NN prediction models were constructed using different gene panel numbers (10, 20, 50, 100, 250) derived from S2N analysis of 21 tumors that were unambiguously classified. The model was then used to predict the membership of the remaining histologically nonclassical specimens. Error rates determine by LOOCV were achieve as low as 14% were reported for a gene panel consisting of 20 genes[2]. Interesting, *vimentin* was reported as a member of the 20 gene panel.

Godard, et at., (2003) conducted cDNA-array analysis of 53 biopsy samples comprising 24 low grade astrocytomas, 8 secondary glioblastomas, and 20 primary Glioblastomas[25]. They demonstrated the application of a novel unsupervised clustering algorithm coupled two-way clustering (CTWC)[90], that finds stable clusters of genes and samples. Clusters that were identified that were able to distinguish recurrent *vs.* primary gliomas. They reported that a cluster comprised of angiogenesis genes could be used to delineate the tumor specimens into primary versus recurrent classes and thus may indicate survival.

1.6 CHAPTER SUMMARY

In summary, microarray technology has evolved beyond simple differential gene expression analysis. It now serves as a platform for multiple types of investigations ranging from sequencing, single nucleotide polymorphism analysis (SNPS), high throughput ligand-DNA interaction screening, and disease diagnosis.

Contemporary knowledge concerning glioma genetics has converged on some important issues. Gene expression patterns can distinguish histological classes of gliomas with an appreciable degree of accuracy. Gene expression in GBMs is highly variable, which may reflect considerable within specimen heterogeneity. Gene expression can be used to reclassify ambiguous tumor specimens more accurately into histological classes that better reflect survival. Several important genes have been consistently identified including genes related to invasion, motility, angiogenesis (*IGFBP*-2), and anti-apoptosis (*DAP*-3). In fact, *IGFBP*s 1 – 7 have been reported differentially expressed in gliomas[2,23,41,86].

In terms of data analysis for microarrays, this body of work has been characterized by increasing sophistication. There have been multiple, disparate, approaches each aimed at extracting particular modes of information. Initially researchers reported simple observations of mean fold changes, which naturally was improved upon by performing statistical tests of significance (*t*-test, ANOVA). The S2N algorithm of Golub, et al., has been employed in many studies seeking to find genes predictive of binary classification[2,23] and clustering has been used extensively.

The following chapters describe the fulfillment of the specific aims outlined in the dissertation proposal and detail the development and validation of our biochip platforms.

These specific aims were: *i)* development, *ii)* production, *iii)* validation of the C3B 10K oligonucleotide microarray (10KO), and *iv)* use of this microarray to conduct a gene expression study aimed at identifying genes predictive of astrocytoma classification. This dissertation tells the story of how our microarray platforms were designed, fabricated, utilized, and convey our novel contributions to the field of biochip engineering. Chapters 2 details the design and development of the 10k human oligonucleotide microarray and makes reference to our published manuscript. Chapter 3 describes the brain tumor class prediction study, including the design and production of our custom spotted 10K human oligonucleotide microarray. It is widely expected that future DNA devices will be indispensable in biotechnology and pharmaceutical research as well as disease diagnosis and prognostic estimation.

CHAPTER 2. DESIGN AND DEVELOPMENT PARAMETERS FOR THE 10K HUMAN OLIGONUCLEOTIDE MICROARRAY

2.0 Design of the Human Oligonucleotide Microarray

The surface chemistry was selected using parameters identified from an initial fabrication experiment[35]. Although microarrays have been fabricated on many types of surfaces[22,31,35,37,91,92,93] we demonstrated that 3-glycidoxypropyltrimethoxysilane (GPS) surface provided higher signal (foreground intensity divided by background) than other common surfaces (Figure 2.1).
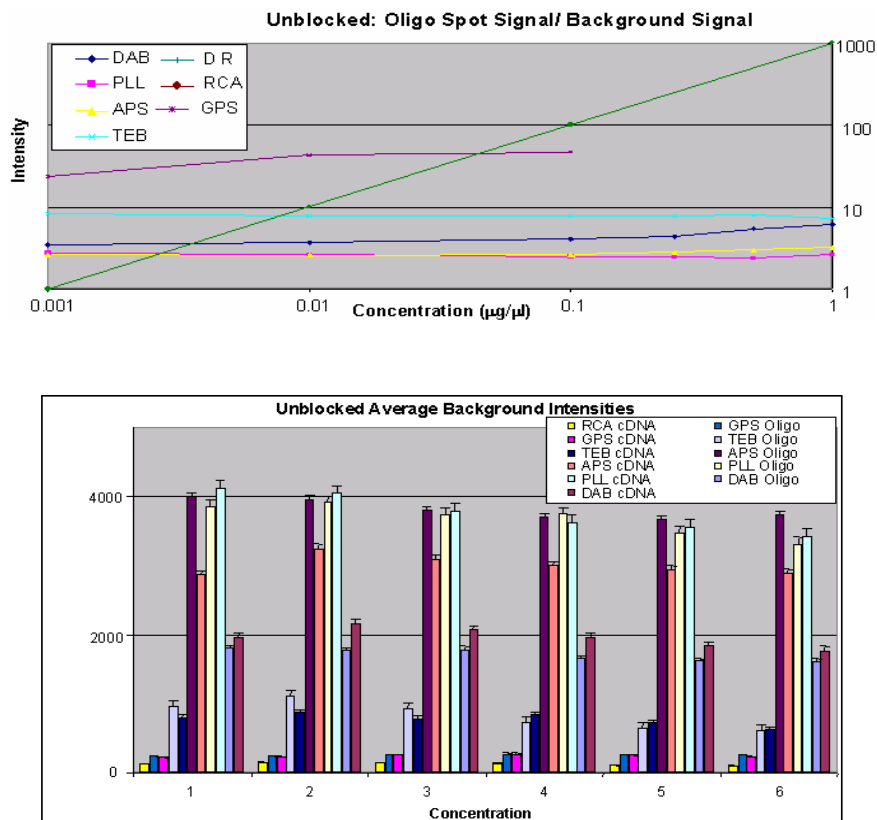
Figure 2.1 Signal (intensity divided by background) of oligonucleotide DNA spotted on five surfaces. The top panel is a line graph showing signal as a function of spotting concentration. The bottom panel is a bar graph with error bars for standard deviation. The data shows that the epoxy silane surface gave the highest signal while spotting concentration leveled off after 0.01 μg/μl. Surface abbreviations: DAB; amino dendrimer, PLL; poly-L-lysine, APS; γ−aminopropaltrimethoxysilane, TEB; tris-EDTA buffer, GPS; glycidoxypropaltrimethoxysilane. (Figure taken from[35])

The GPS surface used in conjunction with amine modified oligonucleotide probes represented a method for covalent attachment of the oligos to the surface [22,136,91]. Because the pH of the spotting solution was determined to be 5.2, we suggest the covalent attachment of the oligonucleotide to the GPS was considered to proceed through an acid catalyzed epoxide ring opening reaction. A simple schematic of the reaction of an amine-

Figure 2.2. Epoxide ring opening reaction and covalent bond formation. Amine terminated oligonucleotide, denoted by R, is covalently attached to the glycidoxysilane molecule.
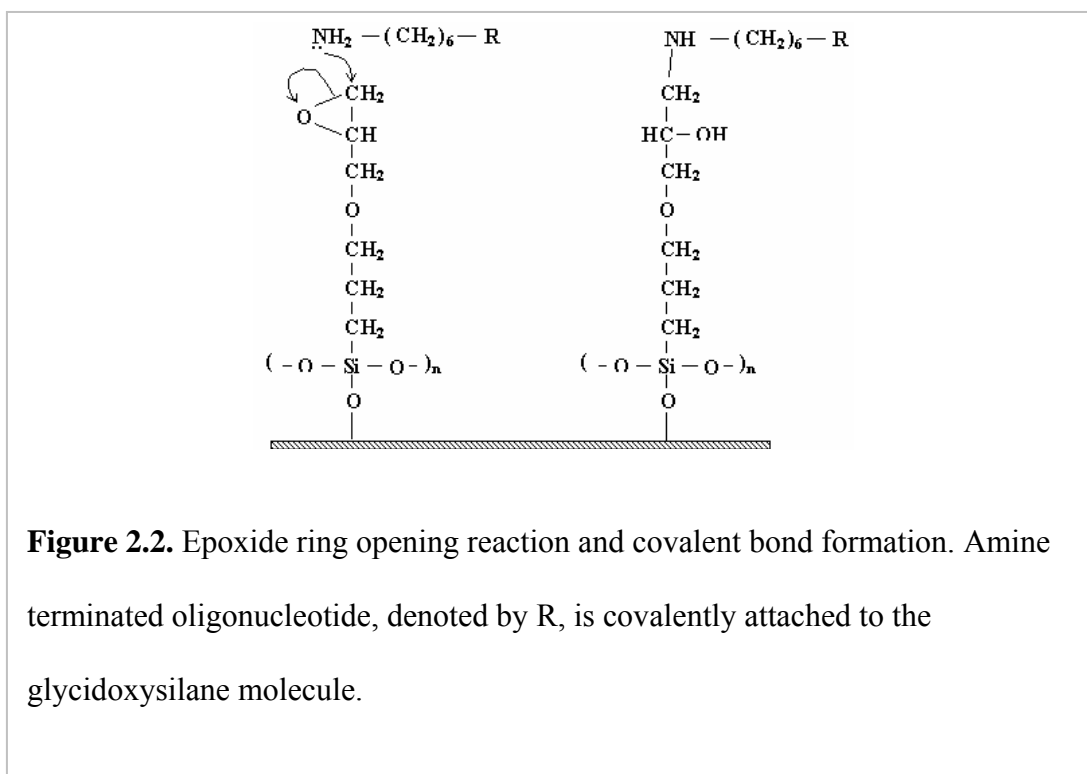
modified oligo with the epoxide ring of a GPS molecule is depicted in Figure 2.2. Thus the GPS surface was chosen as the substrate for immobilization of the C3B human oligonucleotide library to 1in x 3in Goldseal (Cat# 3010, Gold Seal Products) microscope slides. Spotting was preformed using a Cartesian PixSys 5500 microarrayer.

The 10k human oligonucleotide microarray was designed using the MWG 10kA human oligonucleotide library (Cat # 2190-000000, MWG) as the base gene library. Seventeen additional "housekeeping" gene-probes and eleven probes that are also found on the Affymetrix Hu133A chip, listed in appendix A2 (Table A2), were added to the 9,984 5'-C6-amine-terminated and HPLC purified 50-mer oligonucleotides in the MWG set. These additional gene-probes served as internal control features and for future inter-platform data comparisons; an on-going project of the C3B. Seventy-eight additional gene-probes, which were identified though a literature search as being relevant to glioma genetics, were also added to the MWG library. These probes were also purchased from MWG and are listed in appendix A3 and these features are depicted in Figure 2.3.
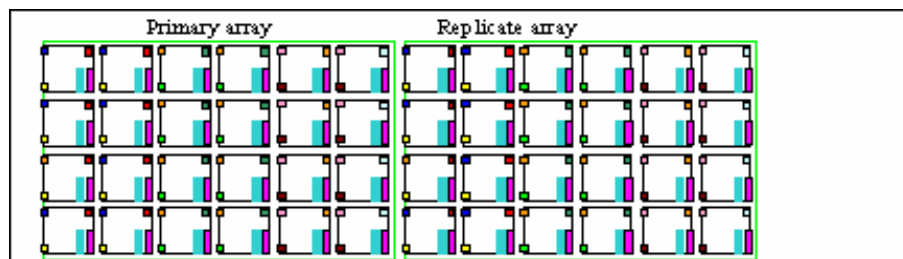
Figure 2.3. The 10k human oligonucleotide microarray. Housekeeping genes are denoted by colored boxes in the corners of the sub-grids. Nonspecific hybridization controls are denoted by aqua rectangles. 3'/5' and bacterial "spiked-in" controls are denoted by magenta rectangles. The features are colored to show location and are not drawn to scale.

2.1 Fabrication methods

To functionalize the surface of the microscope slides, the slides were first solvent cleaned in isopropanol for 1 min at 56°C followed by 1 min in acetone at 56°C. Subsequently the arrays were dried in an Eppendorf 5804 refrigerated centrifuge by spinning for 3 mins at room temperature then placed in a UV ozone cleaner (Model 135500, Boekel) for 10 min. The slides were then sonicated (Model 1510, Branson) in isopropanol at room temperature for 1 min, rinsed in flowing ultrapure water, followed by immersion in RCA (5:1:1, diH20: hydrogen peroxide: ammonium hydrioxide) solution at 60°C for 1 min, rinsed again in ultra-pure water, and dried by centrifugation. The cleaned, dried slides were placed in a 0.1% v/v solution of toluene and GPS for surface modification at 40°C for 30 min. After this incubation, the slides were washed in anhydrous toluene, and cured at room temperature for 48 hrs.

The printing script, executed by the Cartesian software, was custom written in house exclusively for the production of the 10KO such that the base library, including the 78 supplemental oligos was printed first followed by placement of the control features.

Contact printing was performed under 50 % relative humidity using eight silicon spotting quills (Parallel Synthesis). Oligos were spotted at 25 mM in a spotting buffer (pH = 5.2) of 0.75 M betaine and 1.5 X SSC as reported in Diehl et al.. The spotting concentration was specified by the MWG protocol as 50 mM, but was reduced to the 25 mM concentration on the basis of information gleaned from the initial microarray fabrication experiment[35] (Figure 2.1). The primary array was printed in duplicate on each slide yielding 21,168 total features. The primary and replicate sub-arrays were divided into 4 x 12 (48) sub-grids of 21 x 21 (441) spots each (Figure 2.3).

2.2 Quality control features

The quality control features enabled quality prescreening of the hybridized
microarrays. These features were analyzed using the NQC R script written specifically
for the 10KO. At the time the 10KO was designed, the housekeeping genes were thought
to be useful for normalization of microarray data, this notion has been largely
debunked[29,56]. Currently, these probes primarily serve to align the quantification grid of
the Quantarray software.

The probes for the bacterial genes *BioB*, *BioC*, *PheB*, and *ThrC* (Table A2), are
implemented for positive control and uniformity of the hybridization event. They are
spotted at a concentration of 25 mM in row 21, columns 1-4 of each sub-grid, and Cy3
(MWG) labeled complimentary targets were spiked into the hybridization solution at the
final concentrations of 500pM, 250pM, 125pM, and 75pM respectively such that it was
expected to observe a linearly decreasing foreground intensity from these spots. To test
for this, a lack-of-fit test[61] was applied to the Log_2 intensity values of these spots. A p-
value < 0.01 was considered as evidence against the appropriateness of the regression
model i.e., lack of fit.

The BioD series of 10 spots, printed at concentrations in a two fold dilution
starting at 200mM were used to obtain a value for data filtering based on the linear
regression at the intensity on the y-axis corresponding to the graphical intersection at
25mM (Figure 2.4). This information was used to remove spots that displayed intensity at
the non-specific hybridization intensity threshold.
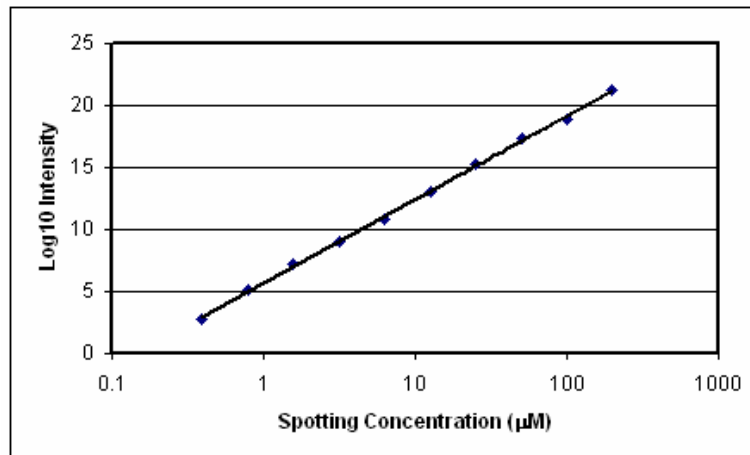
Figure 2.4 Example graph for linear regression of BioD spots. The intersection

of Log_2 intensity and spotting concentration (25μM) was used to determine the

value at which to filter the data for nonspecific hybridization.

Finally, there were six spots that probe for 3' and 5' ends of the transcripts for

GAPDH, ISGF, and Beta Actin. These spots report the fidelity of the RT reaction by

indicating that the distal (5') ends of mRNA transcript are copied with the same

abundance as the 3' end. For these spots the I3'/I5' = IR is $0 \leq IR \leq 3.0$, where I is the

foreground intensity from the spot. Arrays with poor IRs are typically discarded from

further analysis.

CHAPTER 3. MALIGNANCY GRADE AND OUTCOME PREDICTION IN HUMAN GLIOMAS BY DNA MICROARRAY ANALYSIS

ABSTRACT

We report the identification of predictive gene panels for tumor grade and

survival by microarray analysis of 64 glioma samples including WHO grades I, II, and IV

for astrocytomas, and grades II and III for oligodendrogliomas. We demonstrate that

transcriptomic profiles are able to distinguish tumor grade and predict survival class in a

comprehensive set of human gliomas. Prediction analysis of microarrays (PAM)

identified a 22 gene panel capable of distinguishing glioblastoma multiforme (GM)

tumors from oligodendrogliomas (OL) SAM censored survival followed by k-nearest

neighbors (k-NN) class prediction revealed simple survival classes for gliomas. Analysis

using PAM identified 22 genes with a false discovery rate (FDR) = 0.0 that achieved 94%

classification accuracy among GMs and OLs. Survival analysis followed by k-NN class

prediction achieved ~84% classification accuracy for a 3yr simple survival rule with a 50

gene panel, FDR = 0.24. This study supports the use of microarrays in molecular

diagnosis and prognostic estimation for human gliomas for the purposes of patient counseling and treatment planning.

3.0 INTRODUCTION

The instantaneous transcript abundance profile data captured by the DNA microarray has been shown to accurately predict histological class and survival in multiple cancers[2,8,50,62,64,65]. We applied this technology to investigate gene expression in a comprehensive set of human gliomas. Gliomas are a devastating form of brain cancer, leading to >17,000 deaths per year in the United States[80]. Patients diagnosed with glioblastoma multiforme (GM) have a mean survival of ~52 weeks[13,94] and brain cancers are one of the leading causes of death in children[80]. Histologically ambiguous explants are associated with a 30% misclassification rate[2,82,95], and current treatments for gliomas have failed to significantly increase quality of life or survival for the past 25 years[13].

Efforts to understand the molecular etiology of this disease have been facilitated by use of DNA microarrays[6,4,15,17,20,23,41,48]. Initial work identified differential gene expression by a mean Log_2 fold-change (FC) threshold (i.e., FC = 2). This revealed genes such as *ISGFBP-2*, *laminin-8,9 SPARC*, *TIMP-1,2,* c-myc, *vimentin*, *VEGF, PDGFR,* and *TGF-β, and ISGFBP-1,2,5,6* to be differentially expressed in gliomas relative to normal brain[15,20,48]. More sophisticated analysis has revealed distinct patterns of gene expression in gliomas, high variability in GMs, and demonstrated clusters of genes involved in angiogenesis, cell motility, and progression[6,,41,67,86]. Class prediction and survival analysis has revealed gene panels that are capable of reclassifying previously ambiguous gliomas into a more appropriate survival class. Novel clustering methods have identified gene

expression differences in primary *vs.* recurrent gliomas which may indicate therapeutic targets for treatment[2,41,65].

A major future application of DNA microarray technology is the development of a targeted, low density, DNA based diagnostic/prognostic devices. Development of such devices will depend on delineation of an adequate panel of genes that relate to molecular subtype for diagnostics, and outcomes for prognosis[96]. We seek to identify such gene panels for diagnosis of malignancy and prognosis of glioma patients through microarray analysis of glioma samples.

In the current investigation, 62 human glioma samples representing 5 WHO malignancy grades were analyzed using the custom spotted C3B 10K oligonucleotide microarray (10KO). To identify gene panels that were predictive of malignancy grade and survival, feature selection, based on gene expression values, was preformed using methods implemented in prediction analysis for microarrays (PAM)[72] (malignancy grade) and univariate Cox proportional hazards modeling (survival) using censored survival data. Class prediction models were built for malignancy grade (shrunken centroids)[72] and simple survival rules (*k*-NN). It is broadly anticipated that transcriptomic information about disease etiology will enable the development of personalized treatments and drastically improved therapeutic quality[2,25,61,67].

3.1 METHODS AND MATERIALS

3.1.1 Sample acquisition

[‡]Tumor tissue was prospectively collected in the operating room in accordance with VCU IRB-approved protocols (VCU IRB#3031). Samples were snap-frozen in liquid nitrogen within 5 minutes of excision and stored at -86ºC until ready for sectioning and extraction. IRB approved glioma samples were acquired from the VCU medical center campus Broaddus/ Filmore tumor bank and transported to the C3B laboratory on dry ice. Received samples had associated sample ID, histopathological category, and mortality time/ time to censor. For total RNA extraction, tissues were pulverized to a fine powder in pre-cooled nuclease-free mortar and pestle, and pulverized tissue placed directly in TRIzol reagent (Invitrogen, 15596-026) and processed according to manufacturers specifications. Total RNA extraction was followed by clean up by RNeasy columns (Qiagen Inc., Valencia, CA) according to the manufacturer's protocols. Samples were stored at -80°C until removed for reverse transcription.

The quality of the total RNA sample was assessed using a 2100 Bioanalyzer and RNA 6000 LabChips[®] (Agilent, Palo Alto, CA) such that the measured 28s to18s ribosomal quantity ratio ≥ 1.1. The sample set was comprised of six histopathological classes: 25 glioblastomas (GM), 9 anaplastic astrocytomas (AA), 10 pliocytic astrocytomas (PA), 10 oligodendrogliomas (OL), and 10 anaplastic oligodendrogliomas (AO). These samples are listed in appendix 5.

[‡] The work described in this paragraph was preformed by our collaborators: Dr. Tim Van Meter, and the Dr. William Broaddus and Helen Filmore research group.

3.1.2 Experimental design

The hypothesis that a subpopulation of genes within the C3B gene library can be used delineate glioma stage, grade, and patient outcome (survival) was tested using data derived from a standard reference design (Figure 1.3). This design gives several advantages: it allows for open ended data collection (extensible), it enables statistical estimation of all main effects and higher order interactions with the same precision, it allows for large numbers of samples without increasing the complexity of analysis, and down stream clustering analysis is greatly simplified over designs such as a loop design[45,67,97]. This design can also be thought of as a randomized block design[99] where each array represents a block, and arrays are randomized with respect to sample.

For this design, a reference sample ($r = 1$) and a tumor sample ($t = 62$) are co-hybridized to each array such that each array is interrogated by the same reference but a different tumor sample. Stratagene Human Reference total RNA was used as reference total-RNA and was labeled with Alexafluor 647. Each tumor sample was labeled with Alexafluor 555.

3.1.3 Reverse transcription, array hybridization, and labeling

Tumor total-RNA was reverse transcribed according to standard protocols. The Genisphere labeling kit (Genisphere Cat # H500100 and H500110) was used to fluorescently label the reverse transcribed and hybridized oligonucleotide targets. This labeling method utilizes a two-step approach. The first step was the hybridization (for 16 hr) of the reverse transcription (RT) product to the oligonucleotide probes on the surface of the array. The RT product, was synthesized using Genisphere primers that contain a

linker region that binds to the labeling dendrimer in the second hybridization step (4 hr). The labeling dendrimer consisted of a third generation DNA dendrimer that contained ~950 Alexafluor molecules, and the relatively large amount of fluorescent molecules on the dendrimer molecule allowed for the use of extremely small amounts starting material. For the Genisphere labeling method, 0.5 – 2.0 µg/µl of total-RNA are recommended and by comparison, alternative labeling methods such as aminoallyl or dyeconjugated nucleotide labeling require 15 –25 µg/µl of total-RNA. Thus, the Genisphere reagents were chosen for this experiment chiefly because of limited tumor sample total-RNA availability. It has been noted that earlier generation Genisphere products have been associated with a more limited ability to detect fold changes (sensitivity) compared to aminoallyl or dye conjugated nucleotides, however, the current generation products used in this experiment have not been evaluated for their sensitivity.

Stratagene Human Reference total-RNA (Cat # 740000) was chosen as the reference RNA used to hybridize against the sample cDNA and was always labeled with Alexafluor 647. The reference cDNA was reversed transcribed in paired-with-sample10µL reactions then pooled before aliquoting into microcentrifuge tubes containing the tumor RT product. A total of 1µg of tumor total-RNA and reference total-RNA was used for each RT reaction. The RT reactions were preformed in a Scigene heat block with heated bonnet at 52°C using Superscript III (Invitrogen, Cat# 18080-044) reverse transcriptase in a volume of 10µL. Hydrolysis of RNA and neutralization were preformed according to the Genisphere protocol.

Hybridization and labeling took place in Telechem hybridization cassettes using Lifterslips (Erie Scientific, Cat# 25x60I-2-4789). Sealed cassettes were incubated for 16 hrs in an oven at 52^0C for the cDNA hybridization step. The hybridization buffer consisted of 30µl of 2X Enhanced hybridization buffer (Genispere Cat # CW31200S25), 14 µl each of sample and reference RT product solution, in a total volume of 60 µl. Spiked-in probes (2µl of 100X Cy3 labeled oligonucleotides) complementary to the control features *BioB, BioC*, *ThrC*, and PheB was included in the hybridization solution at 500pM, 250pM, 125pM, 67.25pM respectively. The cDNA hybridization solution was pipetted under a Lifterslip that was placed over the array and wrapped in parafilm prior to preparation of the hybridization solution. Post hybridization, arrays were washed for 10 min in medium stringency wash buffer (2X SSC and 0.1% SDS) at room temperature then rinsed 10 times in 2X SSC buffer according to the Genisphere protocol. The arrays were then dried in an eppendorf centrifuge at 1300 RPM for 3 minutes.

Dried arrays were covered with a Lifterslip, wrapped in parafilm and placed back into the hybridization cassette for the labeling step (3DNA hybridization). The 3DNA labeling solution consisted of 30 µl of 2X SDS based hybridization, 3 µl of Alexafluor 555 labeling reagent (Genisphere kit component), 3 µl of Alexafluor 647 labeling reagent (Genisphere kit component), 1 µl of anti-fade reagent, and 23 µl ultrapure water. The labeling solution was pipetted under a Lifterslip and the sealed hybridization cassette was placed in a lab oven 52°C for 4 hours. Washing and drying procedures were carried out in the same manner as in the cDNA hybridization with the exception that the initial wash was 5 mins with 100 µM dithiothreitol (DTT) (Invitrogen, Cat # Y00147) added to the

wash buffers to protect against Alexafluor 647 dye fading (Figure 3.1). All washes were preformed according to Genisphere protocols.

3.1.4 Image acquisition and quantification

Hybridization times were staggered such that arrays completed their labeling hybridization period in 15 min intervals. This was done to minimize AlexaFluor 647 dye fading[99] (Figure 3.1) and arrays were scanned immediately after drying by centrifugation.

Hybridized microarrays were scanned in a ScanArray Express microarray scanner under 90% laser power, 80% PMT gain and 10μm scan resolution. Scanned images were saved as *.tif files for image quantification. Acquired images were quantified using the

Figure 3.1. Scatter plots of AlexaFluor 647 dye fading. Intensity data from the AlexaFluor 647 channel from the initial scan plotted against data from the same array scanned after 10 minutes (left panel), and after 30 minutes (right panel) of dark storage under ambient conditions. The initial scan data were plotted on the x-axis and the subsequent scan data were plotted on the y-axis. Notice, as time progressed, the intensity distribution shifted toward the initial scan data. Red dots indicate a two-fold difference.

QuantArray software from Perkin Elmer. The adaptive circle method of foreground intensity pixel segmentation was used to define the margins of the spots and spots were located using the nominal location feature, i.e., spots were not located using the

automatic algorithm. The resulting quantification output was saved as ANSI tab delimited text files.

3.2 ANALYTICAL METHODS

3.2.1 Quality control and data normalization

Prior to the analysis of the raw data files were modified in two ways, *i*) the text files were deconstructed such that two files were created from each initial data file (decon), and *ii*) the primary and replicate gene intensity values were averaged (gene-averaged). Recall that the 10KO contained a primary and secondary array yielding two measurements for each gene (Figure 2.3), 10584 measurements from the primary array, and 10584 measurements from the replicate array. For the deconstructed (decon) files, this translates into two columns (vectors) in the expression matrix X_{decon} for each sample. This essentially increases the number of measurements by two, and was done to increase the sensitivity of the feature selection for the survival analysis. This paradigm was chosen based on statistical theory[33]. Malignancy classification was preformed on gene-averaged data.

Pre-analytical data treatment proceeded in four steps, two quality control steps and two normalization steps. First, the data were prescreened by NQC (2.3) to detect arrays that performed poorly. This data is given in Appendix 4. Correlation among all pairs of arrays was calculated on a per channel basis, this data is given in Table A6 in (Appendix 6). The correlation metric used was,

$$\rho_{x,y} = \frac{\text{cov}(X,Y)}{\sigma_X \cdot \sigma_y} \qquad [3.1]$$

where σ = the sample standard deviation.

Prior to normalization the $\log_2$ ratio of sample / reference was calculated which serves to control for differences in immobilized probe concentration and spatial effects. The data were then normalized in two steps. First arrays were regularized by standard deviation[45]. This normalization is preformed based on the assumption that all spots within each subgrid on an individual microarray and all spots within each microarray in a set of microarrays should have the same standard deviation for $\log_2 I_{is}/ I_{ir}$, where I is the measured intensity for spot i in the sample channel (s) and the reference channel (r).
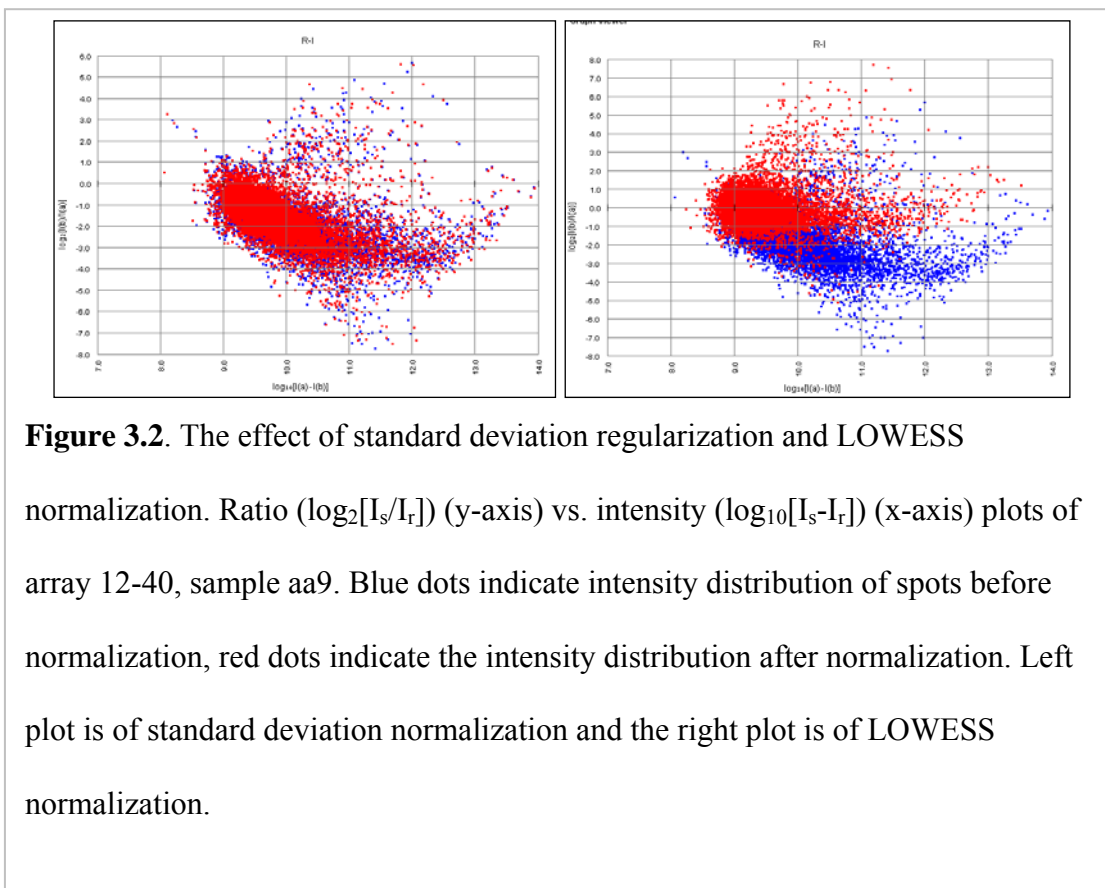
Figure 3.2. The effect of standard deviation regularization and LOWESS normalization. Ratio ($\log_2[I_s/I_r]$) (y-axis) vs. intensity ($\log_{10}[I_s-I_r]$) (x-axis) plots of array 12-40, sample aa9. Blue dots indicate intensity distribution of spots before normalization, red dots indicate the intensity distribution after normalization. Left plot is of standard deviation normalization and the right plot is of LOWESS normalization.

Standard deviation regularization scales the sample and reference channel intensities for

each spot such that the spots within each sub-grid or all spots within each microarray in an experimental set will have the same standard deviation for $\log_2 I_{is}/I_{ir}$. This adjustment was applied to sub-grids within an array to adjust for uneven hybridization and between arrays to normalize for array-to-array differences (i.e., production lot).

Finally, the data were normalized by sub-grid LOWESS normalization[32] with $f =$ 0.4. This normalization was applied to correct curvature in the ratio vs. intensity (RI) plot (Figure 3.2), and to shift the intensity distribution to center around zero on the RI plot. Curvature in the RI plot indicates a dependency of $\log_2 I_{is}/I_{ir}$ ratio distribution on the measured intensity value. Shifting the intensity distribution toward zero on the RI plot helps remove the dye dependent intensity bias of the distribution. Data normalized in this way were used in all subsequent analyses. These normalizations were preformed using the MIDAS program[45].

3.2.2 Feature selection and class prediction

Feature selection is a general term used here to indicate the method by which features (genes) are selected for further analysis. In the context of microarray data, many genes provide little information about a particular biological condition so it is necessary to develop a means for selecting genes that are relevant to the question at hand. Some of these methods previously employed by the microarray community are described in chapter 1. The methods used for the study described in this chapter are PAM, (for malignancy grade), SAM Cox proportional hazards modeling followed by k-NN classification (for survival).

3.2.2.1 Prediction analysis for microarrays

Prediction analysis for microrrays performs automatic a feature selection via shrinking the class specific centroid. This removes genes that fall within the limits of the threshold parameter (Δ)[72] which controls the amount of shrinkage. Classification is preformed by calling unknown sample x_n^* a member of class i by computing the distance to each class specific centroid and giving membership to the nearest centroid. For convenience, the method described below is taken from Tibshirani et al. For a more verbose treatment refer to the manuscript[72].

Briefly, the centroid for gene expression data can be calculated for each class and

Figure 3.3. Illustration of class-wise gene expression centroid. Each class has a centroid based on gene expression values. The values of Δ control the amount of centroid shrinkage. Here, two values of Δ are depicted, and for Δ_2 (in class 3) the amount of shrinkage excludes gene 2 from consideration. This removal is the feature selection step.
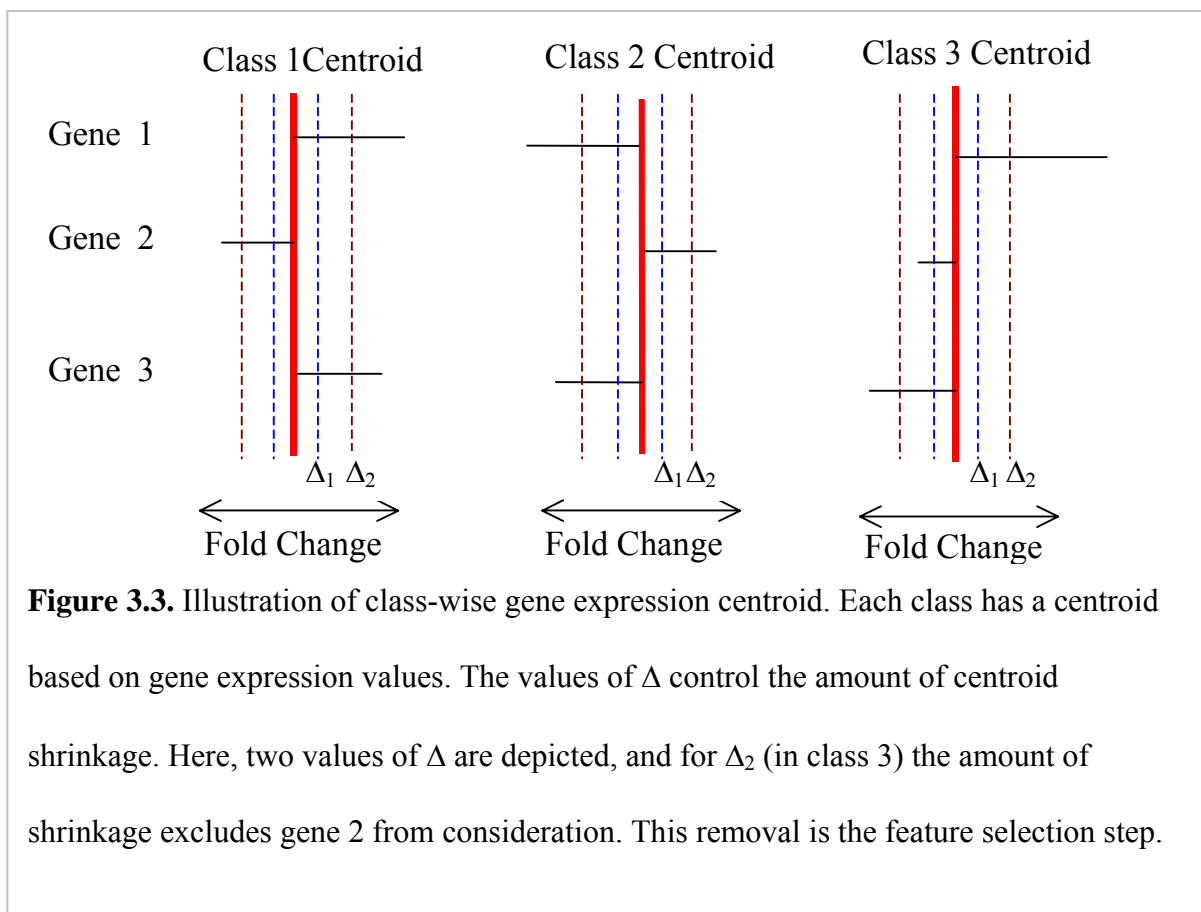
is illustrated in Figure 3.3. If x_{gs} is the expression value for the gth gene where $g =$

$1, 2, \ldots, l$ and s^{th} sample where $s = 1, 2, \ldots, n$, and there are $1, 2, \ldots, c$ classes such that the class labels are denoted by i_c for the n samples in class i, then the gth component of the centroid for class i is $\bar{x}_{gi} = \Sigma_{s \in Ci} x_{gs} / n_i$.

Given

$$d_{gi} = \frac{\bar{x}_{gi} - \bar{x}_g}{m_k \cdot (q_g - q_o)} \quad [\,3.2]$$

where

$$q_g^2 = \frac{1}{s - i} \sum_i \sum_{s \in C_i} (x_{gs} - \bar{x}_{gs})^2 \; [\,3.3]$$

and $m_i = \sqrt{1/s_i + 1/s}$ makes $m_i \cdot q_g$ equal to the estimated standard error of the numerator in d_{gi}. The value for q_o is set such that it is equal to the median value of q_g over the set of genes to remove large values of d_{gi} that arise from chance by low expression levels as in the SAM algorithm[39]. Thus, d_{gi} is essentially a t-statistic and the PAM method shrinks each d_{gi} toward zero (by specifying the Δ parameter) giving d'_{gi} and yielding the shrunken centriods. This can be expressed as,

$$\bar{x}'_{gi} = \bar{x}_g + m_i(q_g + q_o)d'_{gi.} \qquad [3.4]$$

This method of shrinkage can be described as soft thresholding and is described in the following equation,

$$d'_{gi} = sign(d_{gi})(|d_{gi}| - \Delta)_+ \qquad [3.5]$$

where the + indicates to take the positive part ($t_+ = t$ if $t>0$ and zero otherwise)[72]. It can be seen in Eq. 3.5 and Figure 3.3 that the value of Δ sets the threshold for the amount of centroid shrinkage.

The ith discriminant score δ_i is calculated relative to the ith shrunken centroid and is corrected by the class prior probability[72]. The classification rule specifies the class membership choosing the discriminant score that minimizes the distance of the gth test observation to the shrunken centroid.

The latter is a type of classification algorithm that is similar to k-NN with the exception that class membership is determined as a function of the class specific centroid, rather than the k nearest distances to individual class members.

3.1.2.2 SAM censored survival

Survival analysis is a class of methods used for time-to-event or failure time analysis. In engineering these methods are used to estimate such quantities as product life expectancies. In the medical field, they are used to estimate survival time of individuals given data such as stage of disease. The survivor function at time t is the proportion of units in the population for whom $T > t$. For instance, the proportion of individuals still alive at age $t = 95$ years. This function can be denoted $S(t)$

$$S(t) = P(T>t) = 1\text{-}P(T{\leq}t) \qquad [3.6]$$

The most widely used survival model is the Cox proportional hazards model[100,101] whose general form is given in Eq. 3.7.

$$h(t) = [h_0(t)]\, e^{(\beta X)} \qquad [3.7]$$

or equivalently

$$h(t) = [h_0(t)]\, \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n) \qquad [3.8]$$

This is a semi-parametric exponential regression function for the hazard at time *t*, *h(t)*.

The value of the *ßs* are estimated using partial likelihood. It can be seen from Eq. 3.7 that when $X = 0$ the hazard *h(t)* equals the baseline hazard $h_0(t)$. For a single dichotomous independent variable such as 0 for a censored observation and 1 for death, dividing each side by $h_0(t)$ gives the hazard ratio, which indicates the expected change in the risk of the event when X_1 changes from 0 to 1.

The Cox model has been conveniently implemented in the TIGR MeV software SAM module where the SAM distance metric *d(g)* of Eq. 1.14 for the Cox model is

$$d_0(g) = h(t)_g = [h_0(t)_g] \exp(\beta_{1g} X_{1g}) \qquad [3.9]$$

This software accepts as input time-to-death, time-to-censor and gene intensity x_{ig} and finds genes that are significantly related to survival according to the value of Δ selected. Once a list of genes significantly related to survival is determined, any value of *t* (days) can be selected for defining a classification rule. Ideally, we would like to find the minimum number of genes that achieves the lowest classification error rate.

We chose four biologically arbitrary classification rules (i.e. +/-1 yr survival) to select a list of genes that could classify samples with a low error rate. This, of course, throws out information but is nevertheless useful for giving glioma patients and clinicians information for personal and treatment planning. The *k*-NN classifier that was used to test the ability of the genes selected using the survival model was described in detail in section 1.2.3.3, and its error rate was estimated using LOOCV.

3.1.3 Analysis procedure for malignancy grade

Gene-averaged files were formatted for input into the PAM software package. This package operates in the R environment[103] and allows estimation of generalization (classification) error (ε) by v-fold cross validation, for all analyses v = 10.

The v-fold procedure involves partitioning the sample set in to v fractions, calculating the classification error rate with v − 1 fractions, as a function of the threshold parameter Δ, and repeating this procedure for all v fractions. The error rates are averaged and the variance, due to the differences in error estimates among the v iterations, is also calculated.

Plots for cross validation error, and the false discovery rate (FDR) were constructed, and gene lists were generated for selected values of Δ (Figure 3.5). The 22 gene list of GM *vs*. OLs was reported in the results section and was the gene list associated with the lowest FDR and ε.

Analysis was initially conducted on all sample types, and then preformed for selected pairings. The pair-wise comparisons were preformed to select genes from commonly misclassified gliomas. For instance, a histopathologist is unlikely to confuse a PA from a GM, but an AA to GM comparison is more prone to subjective diagnostic error. The specific pair-wise comparisons were GM *vs*. AA, GM *vs*. AO, and GM *vs*. OL.

3.1.4 Analysis procedure for survival

The MeV software package[45] was used for survival analysis. Survival was modeled using only 80% (49) of the samples, the samples not included are listed in Table 3.1. Class prediction models were built using a *k*-nearest neighbors (*k*-NN) *classifier*

using 100% (62) of the available samples. Several classification rules (i.e., survival times: +/- 365 days, 740 days, 1080 days) were applied to genes panels of 100, 50, 20, and 10 genes. Genes were selected for panel inclusion as ranked by the value of d_o (Eq. 4.9). This was done to determine if a reduced suite of genes, fewer than detected significant could be used to build a class predictor. The reduced gene panels were evaluated according to their LOOCV classification error.

This analysis was conducted twice, once with gene-averaged files, and a second time with decon files. This approach was adopted due to the high observed FDR rate after

Table 3.1. Samples not included in survival analysis.

Index	Sample Histology[A]	Survival Time[B]
1	AA	69 d
2	AA	735 c
3	AO	83 d
4	AO	1739 c
5	GM	70 d
6	GM	1103 c
7	GM	421 d
8	OL	1511 d
9	OL	300 d
10	PA	1008 c
11	PA	2402 c
12	PA	1851 c

[A] Sample histology, AA = anaplastic astrocytoma, AO = anaplastic oligodendroglioma, GM= glioblastoma multiforme, OL = oligodendroglioma, and PA = pilocytic astrocytoma.

[B]Survival time is reported for each sample with letter codes: d = dead , c = censored.

gene-averaged survival analysis (Section 3.1.1) because it was expected that the additional data would increase the sensitivity to detect expression differences that were significantly related to survival. After each model fitting and validation step, the genes common to our data and the SMD data set were used make predictions of class membership on the SMD data set. This was done to check the accuracy of our predictive gene panels against a published data set[104]. We were able to obtain 20 GM samples and associated survival data. Since these data were limited in terms of their survival time distribution, only the +/- 1yr survival rule was tested. The arrays used to assay these tumors consisted of ~40,000 cDNAs, with some gene replications.

3.3 RESULTS

3.3.1 Initial cluster analysis

To determine if sample and gene clusters relating to biological variation in the data could be identified through hierarchical clustering, the data were filtered by low intensity resulting in ~1400 genes, then the top 200 (~ 2% of genes) remaining genes with the highest variance were retained. The intensity filter was applied such that spots with intensity lower than a specified cutoff were removed. The number of genes removed was determined by iteratively filtering genes by intensity then filtering the resulting genes by retaining those with the highest variance, then constructing the cluster map. For most iterations, the resulting clustering result failed to indicate biologically relevant clusters (i.e., by sample class). Only after drastic reduction in the genes by first applying the intensity filter, were relevant clusters obtained. Others have performed clustering on the

bulk (i.e., > 200) of the genes present on their microarrays and obtained results that followed expectation (i.e., clusters formed according to sample class or subsets within class)[20,23,41,104].

For the data presented here, the average linkage method used to produce sample

Figure 3.4. Unsupervised clustering of 200 genes after low intensity and variance filtration (top 200 genes (2%)). Clusters were generated by Pearson correlation and average linkage. Classes are color coded; green = NB, dark blue = GM, soft blue = AA, aqua = PA, Dark red = AO, and pink = OL. Low-grade malignancies clustered predominately on the right, high-grade malignancies in the middle, and the normal brain specimens occupy a cluster on the left.

and gene clusters, and Pearson's correlation was used as the similarity metric. Clusters

formed somewhat according to malignancy grade as shown in Figure 3.4. As expected,

the NB samples clustered together.

3.3.2 Class prediction of malignancy grade

3.3.2.1 Classification of all classes

To determine if gene panels could be identified that could distinguish all five

malignancy grades from one another, the microarray data was analyzed using PAM. As a

Figure 3.5. Misclassification rates and FDR curves for all five glioma classes. The

top left plot shows the average misclassification rate (y-axis) of the 5 glioma classes

with variance bars. The bottom left plot shows individual mis-classification rates as

a function of number of genes and Δ for each of the glioma classes. The number of

genes is given on the top x-axis and Δ is given on the bottom x-axis. The right panel

is a plot of the FDR (y-axis) as a function of the number of genes (bottom x-axis),

and the value Δ (top x-axis). Line colors in the left bottom plot indicate the

following: dark red = GM, green = OL, blue = PA, black = AO, and grey = AA

definition, the *overall error* rate (ε) is the average of each *class-wise* error rate. The *class-wise* error rate is the number of times a member of a particular class was classified into an alternate class. This analysis did not indicate a single suite of genes capable of distinguishing all tumor classes with a low ε (i.e., > 30%).

However, this analysis did provide the first indication that the best classification was achieved between GM and OL tumors. Specimens for PA, AA, and AO were not easily distinguishable. The feature selection step was associated with a high FDR, for instance, at $\Delta \sim 0.9$ (Figure 3.5 top x-axis, right panel), the median FDR was ~ 75% (Figure 3.5 y-axis, right panel), for ~ 1900 genes (Figure 3.5 bottom x-axis, right panel). The lack of sensitivity for feature selection can also be observed in the high

Table 3.2. v-fold-crossvalidation error rates for all classes. For the contents of this Table, ε = 0.5, for Δ = 1.0 and ~2300 genes.

Class	AA[A]	AO[A]	GM[A]	OL[A]	PA[A]	Class-wise ε[B]
AA[A]	5	3	0	1	0	0.44
AO[A]	0	1	4	4	1	0.9
GM[A]	2	2	14	2	4	0.42
OL[A]	0	3	0	5	1	0.44
PA[A]	0	0	4	1	5	0.5
					Overall ε[C]	0.48

[A]Tumor class designations.

[B]The error rate is reported in the far right column.

[C]The over-all error rate is the average of the values for each individual class-wise error rate.

misclassification error rate, which achieved its optimum overall misclassification error, ε = 0.48, for Δ = 1.0 and ~2500 genes. Individual error rates are summarized in Table 3.2,

and it can be seen that the tumor classes most likely to be classified correctly were the GMs and the OLs. The AO tumors were most likely to be misclassified followed by the AAs.

3.3.2.2 Class prediction on selected pair-wise comparisons

In a clinical setting, the most relevant need of improving classification occurs among malignancies that are most likely to be misclassified and also have an appreciable disparity in prognosis. Three such comparisons were defined as GM *vs*. AO (Figure 3.6), GM *vs*. AA (Figure 3.7), and GM *vs*. OL (Figure 3.8). In practice, lower grade malignancies are often given elevated status as a conservative measure to ensure the patient is not denied aggressive treatment[2].

The misclassification and FDR plots for these comparisons are given in figs 3.6 – 3.8, and the individual misclassification rates are given in Table 3.3. It can be seen from this data that the comparisons of GM *vs*. AA, and GM *vs*. AO were not as reliable given

Table 3.3. Individual v-fold-cross validation error rates for selected pair-wise comparisons.

Comparison	[A]Threshold	[B]Overall ε	GM ε	[C]Other Class ε	# Genes	Median FDR
GM *vs*. AA	1.2	0.24	0.08	0.66	~420	0.5
GM *vs*. AO	1.4	0.33	0.22	0.60	~150	0.5
GM *vs*. OL	2.0	0.09	0.13	0.0	~22	0.0

[A]The value of Δ is specified and all other parameters are given as a function of Δ.

[B]The overall ε is the average of the individual ε's from the binary classification.

[C]The "Other Class" column gives the ε for the other member of each comparison (i.e., AA, AO, or OL).

the high FDR and class-wise error rates. The comparison of GM to OL did yield a set of genes that were capable of classifying the tumors with a low ε and was associated with a low FDR. These genes are listed in Table 3.4.

Figure 3.6. Misclassification rates and FDR curves for GM vs AO glioma specimines. The top left plot shows the average misclassification rate (y-axis) of the 5 glioma classes with variance bars. The bottom left plot shows individual mis-classification rates as a function of number of genes and Δ for each of the glioma classes. The number of genes is given on the top x-axis and Δ is given on the bottom x-axis. The right panel is a plot of the FDR (y-axis) as a function of the number of genes (bottom x-axis), and the value Δ (top x-axis). Line colors in the left bottom plot indicate the following: green = GM, and red = AO.

Figure 3.7. Misclassification rates and FDR curves for GM *vs.* AA glioma

specimines. The top left plot shows the average misclassification rate (y-axis) of the

5 glioma classes with variance bars. The bottom left plot shows individual mis-

classification rates as a function of number of genes and Δ for each of the glioma

classes. The number of genes is given on the top x-axis and Δ is given on the bottom

x-axis. The right panel is a plot of the FDR (y-axis) as a function of the number of

genes (bottom x-axis), and the value Δ (top x-axis). Line colors in the left bottom

plot indicate the following: green = GM, and red = AA

Figure 3.8. Misclassification rates and FDR curves for GM *vs*. OL glioma specimines. The top left plot shows the average misclassification rate (y-axis) of the 5 glioma classes with variance bars. The bottom left plot shows individual mis-classification rates as a function of number of genes and Δ for each of the glioma classes. The number of genes is given on the top x-axis and Δ is given on the bottom x-axis. The right panel is a plot of the FDR (y-axis) as a function of the number of genes (bottom x-axis), and the value Δ (top x-axis). Line colors in the left bottom plot indicate the following: red = GM, and green = OL.

These results indicate that the gene expression data collected during this experiment is well suited for predicting class membership among GM and OL tumors.

However, the data does not appear as adequate for predictions of class membership

Table 3.4. Gene panel consisting of 22 genes. Top 22 genes yielding a low v-fold-crossvalidation error rate between GM and OL tumors. Here $\Delta = 2.2$.

	Gene Name
1	NM_003380_1 vimentin VIM
2	NM_000582_1 secreted phosphoprotein 1 osteopontin bone sialoprotein I
3	NM_001553_1 insulin like growth factor binding protein 7 IGFBP7
4	NM_021103_1 thymosin beta 10 TMSB10 mRNA
5	NM_004202_1 thymosin beta 4 Y chromosome TMSB4Y
6	NM_000146_1 hypothetical gene supported by BC002991 NM_000146
7	NM_021025_1 homeo box 11 like 2 TLX3 mRNA
8	NM_004048_1 beta 2 microglobulin B2M
9	NM_001276_1 chitinase 3 like 1 cartilage glycoprotein 39 CHI3L1
10	NM_025126_1 hypothetical protein FLJ21786
11	NM_004203_1 membrane associated tyrosine and threonine specific cdc2 inhibitory kinase
12	NM_025024_1 hypothetical protein FLJ14082
13	NM_016610_1 Toll like receptor 8 LOC51311
14	NM_004355_1 CD74 antigen invariant polypeptide of major histocompatibility complex
15	NM_001444_1 similar to fatty acid binding protein 5 psoriasis associated
16	AB061838 ribosomal protein S3 2
17	Transcription elongation factor A SII 3
18	NM_003254_1 tissue inhibitor of metalloproteinase 1, TIMP1
19	NM_001780_1 CD63 antigen melanoma 1 antigen CD63
20	NM_002045_1 growth associated protein 43 GAP43
21	NM_018601_1 hypothetical protein PRO1446
22	NM_003927_1 methyl CpG binding domain protein 2 MBD2

among the other two comparisons.

3.3.3 Class prediction of patient survival

Genes whose expression related to survival were identified using SAM censored survival available in the TIGR MeV software package. A unique feature of SAM is that it allows the user to select the FDR by setting the value of Δ (the difference between observed and expected values of $d(g)$) according to what the researcher considers tolerable. In addition, the user must select a value for the permutations B, to compute significance. Initially, this analysis was conducted with gene-averaged files with 20% of the samples left out (see Table 5.1). The results of the model fitting step with $B = 250$, indicated an unusually high FDR = ~0.74 for 162 genes, meaning that ~74% of the genes were falsely declared significant.

We temporarily ignored the high FDR and built a k-NN prediction model, with k = 5, to test each gene panel for three dichotomous survival rules (+/- 1yr, 2yr, 3yr). Samples that were censored before the decision cut off were excluded from the LOOCV estimator. For instance if the survival rule was +/- 1yr and a sample was censored at 280 days, it was excluded (Table 3.5). The gene panels (10, 20, 50 and 100 genes) were nevertheless capable of producing LOOCV error rates ranging from 9 to 28 % depending on the gene panel and survival rule. The classifier achieved its optimum $\varepsilon = 0.09$ (9%) for $k = 5$ with the gene panel consisting of 100 genes and the survival rules for +/- 2 and +/- 3 yrs. Earlier survival times were associated with higher values for ε but none were above 28%. One-year survival was associated with 23-28% ε depending on the gene panel. The two-year survival had a slightly broader range for ε (8 -20 %). Three-year survival had a similar profile (ε =8-21%) to the 2yr rule. Error rates for all gene panels

and classification rules are listed in Table 3.5. In general the 50 gene panel preformed the best regardless of the decision rule, while the 10 gene panel preformed the worst.

The genes that resulted from this initial analysis were tested against the SMD data set for their ability to predict a +/- 1year survival rule. It can be seen from Table 4.5 that there was not good agreement between the two data sets with the 38 gene panel predicting +/- 1yr survival with an ε of 47%

The survival analysis was then repeated with the decon files. For $B = 250$, an FDR of 0.24 was achieved for 104 genes. These 104 genes are listed in appendix 7 and the LOOCV results are listed in Table 3.6. The k-NN model was built with 100% of the samples with the exception that censored observations were withheld if the censoring occurred before the survival rule. For this analysis $k = 7$.

Generally, the error rates went up compared to the gene-averaged analysis, with the lowest being 16% for the 3yr survival rule and the 50 gene panel. The 2yr survival rule was associated with the lowest range of error rates and while the 1 and 3 yr rules were associated with the lowest rate. For this analysis, the error rates were closer to the error rates observed by applying common genes to prediction in the SMD data set. This and the observed decrease in the FDR seems to indicate that this data is more accurate than the gene-averaged data even though some of the same genes were called significant in the survival analysis as in the decon analysis. Finally, the performance of a given gene panel seemed to be affected by the survival rule such that a given gene panel did not necessarily perform the same for all survival rules.

Table 3.5. LOOCV error rates as a function of gene panel and survival rule.

Gene Panel	Survival Rule	LOOCV Error %	[A]Class 1 Error %	[A]Class 2 Error %	Sample Total
10	+/- 1yr	17/60 = 28	6/37= 16	11/23=48	23+37=60
20		13/6 = 22	5/37 = 13	8/23=35	
50		14/60 = 23	5/37 =13	9/23=39	
100		15/60 = 25	4/37= 10	11/23=48	
10	+/- 2yr	12/60 = 20	5/31=16	7/29=24	31+29 =60
20		11/60 = 18	4/31=13	7/29=24	
50		12/60 = 20	5/31=16	7/29=24	
100		5/60 = 8	1/31=3	4/29=14	
10	+/- 3yr	12/56 = 21	4/31=13	7/23=30	33 +23 = 56
20		9/56 = 16	3/33=9	6/23=26	
50		8/56 = 14	2/33=6	6/23=26	
100		5/56 = 8	1 /33=3	4 /23=17	
SMD Data					
4 of 10	+/- 1yr	16 / 19 = 84	8/11=73	8/8=100	11 + 8 = 20
7 of 20		12 / 19 = 63	5/11=45	7/8=88	
17 of 50		15 / 19 = 78	7/11=64	8/8=100	
38 of 100		9 / 19 = 47	4 /11=36	5 /8=63	

[A]Class 1 error and class 2 error reflect the number of samples that were assigned to the wrong class. For instance, for the 10 gene panel and +/- 1yr survival rule, 6 of the 37 samples that had less than 1 year survival were misclassified, while 11 of the 23 samples that had greater than 1 year survival were misclassified. This information indicates whether one class is misclassified at a higer rate than the other class. The gene panels for the SMD data were composed of only those genes that were in common between the two array platforms (C3B and SMD).

Table 3.6. LOOCV error rates as a function of gene panel and survival rule.

Gene Panel	Survival Rule	LOOCV Error %	[A]Class 1 Error %	[A]Class 2 Error %	Sample Total
10	+/- 1yr	30/122 = 24	19/44=43	11/78=14	44+78=122
20		35/122= 28	21/44=47	14/78=18	
50		40/122 =32	23/44=52	17/78=22	
100		**29/122 = 18**	13/56=23	16/54=29	
10	+/- 2yr	33/118 = 28	18/62=29	15/56=26	62+56 =118
20		32/118 = 27	17/62=27	15/56=26	
50		27/118 = 23	15/62=24	12/78=15	
100		28/118 = 24	15/62=24	12/78=15	
10	+/- 3yr	35/110 = 32	18/56=32	17/54=31	56+54=110
20		31/110 = 28	15/56=27	16/54=29	
50		**18/110 = 16**	9/56=16	9 /54=17	
100		29/110 = 26	13/56=23	16/54=29	
SMD Data					
2 of 10	+/- 1yr	9/19 = 47 (37)	5/11=45	4/8=50	11+8=19
5 of 20		13/19 = 68 (34)	7/11=63	6/8=75	
19 of 50		8/19 =42 (38)	4/11=36	4/8=50	

[A]Class 1 error and class 2 error reflect the number of samples that were assigned to the wrong class. For instance, for the 10 gene panel and +/- 1yr survival rule, 6 of the 37 samples that had less than 1 year survival were misclassified, while 11 of the 23 samples that had greater than 1 year survival were misclassified. This information indicates whether one class is misclassified at a higer rate than the other class. The gene panels for the SMD data were composed of only those genes that were in common between the two array platforms (C3B and SMD).

3.4 DISCUSSION

Unsupervised clustering revealed that the data derived from this set of hybridizations do not generally form biologically meaningful clusters with out extensive intensity and variance filtering. As a first step in data analysis, clustering can be used to subjectively ascertain the level of noise present in the data. This analysis seems to indicate that there is sufficient noise in the data as to mask the variation strictly due to gene expression. This does not preclude the ability to use tests of significance to uncover differential gene expression, however, it does indicate that low level regulation may be difficult to identify with a high degree of certainty.

The analysis of malignancy grade indicated the data that we collected was not especially capable of predicting class membership among most classes assayed. There may be several reasons for this. It is known that histologically similar tumors (such as GMs *vs.* AAs) are most likely to be histologically misclassified, and total-RNA samples taken from these specimens may have been affected by this likelihood. Further, the AA specimens are grade III malignancies and are highly likely to progress to GM tumors even after treatment[13,14,80] thus a large degree of biological similarity may also partially account for our inability to find substantial differences in gene expression among these tumors. The sample set we assayed was characterized by a large number of classes with a relatively small number of specimens from each class. This too may have limited our ability to detect stark difference among the tumors. Finally, it has been reported that the Genisphere labeling technique is not as efficient at reproducing fold changes as other labeling strategies (i.e., aminoallyl or dye conjugated nucleotides). However the fact that

it is useful when sample amounts are limited (i.e., < 2 µg) dominated our decision when selecting the labeling technique.

However, good feature selection and prediction results were achieved for the comparison of GMs to OLs. This part of the analysis revealed 22 genes that were significantly different between these two classes. Two genes identified, *vimentin* and *CD74,* have been reported elsewhere to be important in distinguishing these two grades of malignancy[2], and because there is such a disparity in the survival between these classes, these genes have been shown to be useful prognostic indicators. Further, *thymosin* β10 has been implicated in anti-angiogenesis, tumor progression, and neuroblastoma development[105-107]. This gene suppresses Ras function, which inhibits angiogenesis and tumor growth[108]. In our study it was unregulated in OLs vs. GMs, indicating that it may be a factor in the slow growing nature of OL neoplasms. These results lend confidence in the accuracy of our findings with respect to classification of GMs *vs.* OLs. Finally this analysis identified *insulin-like growth factor binding protein* 7 (*IGFBP*-7) as being a good distinguishing indicator of GMs and OLs. While many *IGFBP*s have been implicated in glioma genetics, this is the first observation of *IGFBP*-7 as being a factor in diagnosis of the malignancy grade of gliomas.

Survival analysis of decon data revealed 104 genes significantly related to survival with an FDR = 0.24. Genes that were common to the C3B and SMD platforms showed marginal agreement with regard to the 1yr survival rule. This may be due to several factors including small sample size, limited common genes between data sets and gene panels, differences in platform performance, and large numbers of censored

observations in our data set. Therefore follow up on these subjects may prove useful in a subsequent analysis. The data was internally consistent, meaning that genes found in the 80% of tumors analyzed generalized well to the full 100% of samples we assayed. However, it did not seem to be universally consistent, meaning that it did not predict 1yr survival as well with the SMD data set.

The gene *cysteine rich angiogenic inducer* was identified to be significantly related to survival and was generally over expressed in the GM specimens relative to the lower grade malignancies particularly the PAs. This may indicate the increased level of angiogenesis common to these tumors. This gene has been shown to be involved in recurrence and metastasis in hepatocellular carcinomas[109]. The gene *sarcoma amplified sequence* has been reported to be involved in growth and motility in osteosarcomas[110] and was generally unregulated in the high grade tumors in our data.

Overall the findings of this study seem to have been limited by low sensitivity. Several features of the analysis seemed to indicate this. Initial clustering based on the majority of the genes failed to show significant clusters. Only after drastic reduction in the number of genes were reasonable clusters obtained (Figure 3.4). Calculated FDRs for most analyses were observed to be higher than expected.

Finally, class prediction error rates were also higher than initially expected. With regard to the last point however, it has been noted elsewhere that classification error rates are higher for brain tumor microarray data than for other neoplasms (i.e., lung, leukemia, colon, and prostate)[111], and it has been reported a 29.7% misclassification was observed for brain tumor microarray data suggesting that our data, while it did not agree well with the SMD data set, may not be that far off[111]. It may be that brain tumors represent a

particularly challenging prediction problem and this may stem from the observed high variability of gene expression in gliomas and GMs in particular.

Finally, from a gene expression point of view, the survival rules chosen may seem arbitrary, but from a clinical standpoint they represent important knowledge on how aggressive to be with treatment. For instance, if a histologically classified oligodendroglioma patient has a gene expression profile that predicts less than one-year survival than the clinician would be prudent to counsel a more aggressive therapy than would be typical for this disease. In summary, the data presented did indicate several interesting genes that may serve well as diagnostic and prognostic indicators. It also demonstrated that care must be taken to avoid over optimism as even noisy data can be internally consistent. It is therefore important to design experiments around several corroborating techniques to demonstrate the accuracy of the findings.

3.4 FUTURE WORK

For the investigation described in chapter 3, there are several analyses that can be conducted in the future to make more conclusive arguments with regard to survival. The sensitivity for survival analysis may have been limited by the lack of covariates such as patient age, prior health, or histological class in the model specification. A multivariate Cox model may improve the FDR by accounting for covariates, such as age, that are known to have significant prognostic value in patients with brain tumors. In our analysis, all the PA tumors represented patients that were juvenile, and this particular grade of malignancy is most likely to have a favorable prognosis. Detection of genes that are significantly related to survival would also be improved by a larger sample size. In this

case we ran only 50 (80%) tumors for survival analysis and this may have limited our ability to detect genes related to survival. Even so, the SMD data were effectively modeled without the presence of covariates in the model and achieved a low FDR (data not presented) suggesting that this data was more robust to the effects of covariates.

For the prediction step, the error rates may be more accurately represented if use was made of v-fold cross validation as opposed to leave-one-out cross validation. While both types are known to be unbiased, the latter is associated with a higher variance and thus may be over optimistic[67,70]. Together these suggestions may improve agreement of our data to independent data sets.

Finally, it may be useful to run the same samples on Affymetrix (or other commercial products) arrays to get an independent indication of the correlation among C3B arrays. In addition, a completely independent set of samples could be assayed on both platforms which would be helpful in determining how accurate the results are from each single platform.

REFERENCES

1. Schena, M., Shalon, D., Davis, R.W., Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-470 (1995).

2. Nutt, C., Mani, D., Betensky, R., Tamayo, P., Cairncross, G ., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M., Batchelor, T., Black, P., von Deimling, A., Pomeroy, S., Golub, T., Louis, D. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research* 63, 1602-1607 (2003).

3. Liefers, G.J., Tollenaar, R.A. Cancer genetics and their application to individualized medicine. *Eur J Cancer* 38, 872-879. (2002).

4. Mariani, L., Beaudry, C., McDonough, W., Hoelzinger, D., Kaczmarek, E., Ponce, F., Coons, S., Giese, A., Sieler, R., Berens, M. Death-associated protein 3 (Dap-3) is overexpressed in invasive glioblastoma cells in vivo and in glioma cell lines with induced motility phenotype in vitro. *Clinical Cancer Research* 7, 2480-2489 (2001).

5. Bertucci, F., Salas, S., Eysteries, S., Nasser, V., Finetti, P., Ginestier, C., Charafe-Jauffret, E., Loriod, B., Bachelart, L., Montfort, J., Victorero, G., Viret, F., Ollendorff, V., Fert, V., Giovaninni, M., Delpero, J., Nguyen, C., Viens, P., Monges, G., Birnbaum, D., Houlgatte, R. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene* 23, 1377-1391 (2004).

6. Huang, H., Colella, S., Kurrer, M., Yonekawa, Y., Kleihues, P., Ohgaki, H. Gene expression profiling of low-grade diffuse astrocytomas by cDNA arrays. *Cancer Research* 6869-6874 (2000).

7. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, C., Bloomfield, C., Lander, E. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537 (1999).

8. Beer, D., Kardia, S.,Huang, C.,Giodano, T.,Levin, A.,Misek, D.,Lin, L.,Chen, G.,Gharib, T.,Thomas, D.,Lizyness, M.,Kuick, R.,Hayasaka, S.,Taylor, J., Lannettoni, M.,Orringer, M.,Hanash, S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 8, 8, 816-824 (2002)

9. Lacroix M, Zammatteo N, Remacle J, Leclercq, G. A low-density DNA microarray for analysis of markers in breast cancer. *Int J Biol Markers* 17:5-23., (2002).

10. Perou, C., Sùrlie, T.,Eisen, M.,Rijns, M.,Jeffreyk, S.,Rees, C.,Pollack, J.,Ross, D., Johnsen, H., Akslen, L.,FlugeI, E., Pergamenschikov, A.,Williams, C.,Zhus, S.,Lùnning, P.,Bùrresen-Dale, A.,Brown, P.,Botstein, D. Molecular portraits of human breast tumors. *Nature* (2000).

11. Xiao-jun Ma, R.S., J. Todd Tuggle, Justin Gaudet, Edward Enright, Philip McQuary, Terry Payette, Maria Pistone, Kimberly Stecker, Brian M. Zhang, Yi-Xong Zhou, Heike Varnholt, Barbara Smith, Michelle Gadd, Erica Chatfield, Jessica Kessler, Thomas M. Baer, Mark G. Erlander, and Dennis C. Sgroi Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences* 100, 5974-5979 (2003).

12. Ohmine, K., Ota, J.,Ueda, M.,Ueno, S.,Yoshida, K.,Yamashita, Y.,Kirito, K.,Imagawa, S.,Nakamura, Y., Saito, K.,Akutsu, M., Mitani, K., Kano, Y.,Komatsu, N.,Ozawa, K.,Mano, H. Characterization of stage progression in chronic myeloid leukemia by DNA microarray with purified hematopoietic stem cells. *Oncogene* 20, 8249-8257. (2001).

13. Shapiro, J.R. Genetic alterations associated with adult diffuse astrocytic tumors. *American Journal of Medical Genetics* 115, 194-201 (2002).

14. Giese, R., Bjerkvig, R., Berens, M.E., Westphal, M. Cost of migration: invasion of malignant gliomas and implications for treatment. *Journal of Clinical Oncology* 21, 1624-1636 (2003).

15. Markert, J., Fuller, C., Gillespie,Y., Bubien, J., McLean, L., Hong, R., Lee, K., Gullans, S., Mapstone, T., Benos, D. Differential gene expression profiling in human brain tumors. *Physiological Genomics* 21–33 (2001).

16. Reilly, T., Bourdi, M., Brady, J., Pise-Masison, C., Radonovich, M., George, J., Pohl, L. Expression Profiling of Acetaminophen Liver Toxicity in Mice Using Microarray Technology. *Biochemical and Biophysical Research Communications* 321-328 (2001).

17. Watson, M., Perry, A., Budhjara, V., Hicks, C., Shannon, W., Rich, K. Gene expression profiling with oligonucleotide microarrays distinguishes world health organization grade of oligodendrogliomas. *Cancer Research* 61 1825-1829 (2001).

18. Wang, D., Coscoy, L., Zylberberg, M., Avila, P., Boushey, H., Ganem, D., DeRisi, J. Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences* 99, 15687-15692 (2002).

19. Brown, V., Ossadtch, A., Khan, A., Yee, Simon,. Goran, L., Melega, W., Cherry, S., Leahy, R., Smith, D. Mulitplex three-dimensional brain gene expression mapping in a mouse model of parkeinson's disease. *Genome Research* 12, 868-884 (2002).

20. Sallinen, S.S., P. Haapasalo, H. Helin, H. Helen, P. Schrami, P. Kallioniemi, O. Kononen, J. Identification of Differentially Expressed Genes in Human Gliomas by DNA Microarray and Tissue Chip Techniques. *Cancer Research* 6617- 6622 (2000).

21. Schena, M. Microarray Analysis, Edn. 1. (John Wiley & Sons, New Jersey; 2003).

22. LaForge KS, Shick, V., Spangler, R., Proudnikov, D., Yuferov, V., Lysov, Y., Mirzabekov, A., Kreek, M.J. Detection of single nucleotide polymorphisms of the human mu opioid receptor gene by hybridization or single nucleotide extension on custom oligonucleotide gelpad microchips: potential in studies of addiction. American *Journal of Medical Genetics* 96:604-615, (2000).

23. Shai, R., Shi, T., Kremen, T., Horvath, S., Liau, L., Cloughesy, T., Mischel, P., Nelson, S. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene* 22:4918-4923, (2003).

24. Liang, Y., Diehn, M., Watson, N., Bollen, A., Aldape, K., Nicholas, K., Lamborn, K., Berger, M., Botstein, D., Brown, P., Isreal, M. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proceedings of the National Academy of Sciences* 102:5814-5819, (2005)

25. Godard, S., Getz, G., Delorenzi, M., Farmer, P., Kobayashi, Hiroyuki., Desbaillets, I., Nozaki, M., Discerens. A-C., Hamou, M-F., Dietrich, P-Y., Regli, Luca., Janzer, R., Bucher, P., Stupp, R., Tribolet, N., Domany, E., Hegi, M. Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. *Cancer Research* 63, 6613-6625 (2003).

26. Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D.,

Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., Golub, T. R. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415, 436-442. (2002).

27. Lindroos, K., Liljedahl, U., Raitio M., Syvanen, A. Minisequencing on oligonucleotide microarrays: comparison of immobilization chemistries. *Nucleaic Acids Research* 29, E69 (2001).

28. Kerr, M., Martin, M., Chruchill, G. Analysis of Variance for Gene Expression Microarray Data. *Journal of Computational Biology* 7, 819-837 (2000).

29. Thellin, O., Zorzi, W., Lakaye, B., Borman, B., Counams, B., Hennen, G., Grisar, T., Igout, A., Heinen, E. Housekeeping gene as internal standards: use and limits. *Journal of Biotechnology* 75, 291-295 (1999).

30. Irizarry, R. Hobbs, B., Colin, F., Beazer-Barclay, Y., Antonelleis, K., Scherf, U., Speed, T. Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics* in press (2003).
31. Wong, W.H. in The analysis of gene expression data: Methods and software. (eds. G. Parmigiani, E. Garrett, R. Irizarry & S. Zeger) 120-141 (Springer, New York; 2003).

32. Yang, Y. H., Dudoit, S., Luu, Percy., Lin, D., Peng, V., Ngai, J., Speed, T. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 30, e15 (2002).

33. Pan, W., Lin, J., Le, C. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology* 3 (2002).

34. Shalon, D., Smith, S.J., Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hyridization. *Genome Research* 6, 639-645 (1996).

35. Taylor, S., Smith, S., Windle, B., A., Guiseppi. The impact of surface chemistry and blocking stratagies on DNA microarrays. *Nucleaic Acids Research* (2003).

36. Hegde, P., Qi R., Abernathy K., Gay C., Dharap S., Gaspard R., Earle-Hughes J., Snesrud E., Lee N., Quackenbush J. A concise guide to cDNA microarray analysis. *Biotechniques* 29, 548-562 (2000).

37. Zammatteo, N., Jeanmart, L., Hamels, S., Courtois, S., Louette, P., Hevesi, L., Remacle, J. Comparison between different strategies of covalent attachment of DNA to glass surfaces to build DNA microarrays. *Analytical Biochemistry* 280, 143-150 (2000).

38. Montgomery, D. Design and Analysis of Experiments, Edn. 1st. (John Wiley & Sons, New York; 1976).

39. Tusher, V., Tibshirani, R., Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 98, 5116-5121 (2001).

40. Dudoit, S., Yang, Y., Callow, M., Speed, T. Statistical Methods for Identifying differently Expressed Genes in Replicated cDNA Microarray Experiments. *Statistica Sinnica* 12, 111-139 (2002).

41. Rickman, D. et al. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Research*, 6885-6891 (2001).

42 Kerr, K.,Churchill, G. Experimental design for gene expression microarrays. *Biostatistics* 2, 183-201 (2001).

43. Wolfinger, R.D. et al. Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. *Journal of Computational Biology* 8, 625-637 (2001).

44. Simon, R., Radmacher, M., Dobblin, K. Design of studies using DNA microarrays. *Genetic Epidemiology* 23, 21-36 (2002).

45. Dobbin, K., Simon, R. Comparison of microarray design for class comparison and class discovery. *Bioinformatics* 18, 1438-1445 (2002).

45. Saeed, A. et al. TM4: a free open-source system for microarray data management and analysis. *Biotechniques* 34, 374-378 (2003).

46. Nadon, R., Shoemaker, J. Statistical issues with microarrays: processing and analysis. *Trends in Genetics* 18 (2002).

47. Dudoit, S., Fridlyand, J., Speed, T. Comparison of discrimination methods for the classification of tumors using gene expression data. *Mathematical Sciences Rsearch Institute, Berkeley CA.* Tecnincal report #576 (2000).

48. Ljubimova, J., Lakhert, A., Loksh, A., Yong, W., Riedinger, M., Miner, H., Sorokin, L., Ljubimov, A., Black K. Overexpression of alpha-chain-4-containing laminins in human glial tumors indentified by gene microarray. *Cancer Research* 5601-5610, (2000)

49. Bijlani, R., Cheng, Y., Pearce, D., Brooksand, A., Ogihara, M. Prediction of biologically significant components from microarray data: Independently Consistent Expression Discriminator (ICED).

50. Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., Brown, P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* 1, 0003.0001–0003.0021 (2000).

51. Anthony, R., Brown, T., French, G. DNA array technology and diagnostic microbiology. *Expert Reviews in Molecular Diagnostics* 1, 30-38 (2001).

52. Eickhoff, B., Kron, B., Schick, M., Poustka, A., van der Bosch, J. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Research* 27:e33, (1999)

53. Scheck, A.C., Mehta, B.M., Beikman, M.K., Shapiro, J.R. BCNU-Resistant Human Glioma Cells with Over-Representation of Chromosomes 7 and 22 Demonstrate Increased Copy Number and Expression of Platelet-Derived Growth Factor Genes. *Genes Chromosome Cancer* 8 (1993).

54. Wu, H., Kerr, K., Cui, X., Churchill, G. in The analysis of gene expression data: methods and software. (eds. G. Pargmigiani, E. Garrett, R. Irizarry & S. Zeger) 313-341 (Springer, New York; 2002).

55. Tseng, G., Oh, M., Rohlin, L., Liao, J., Wong, W. Issues in cDNA microarray analysis: quality filtering channel normalization, models of variations and assessment of gene effects. *Nuclei Acids Research* 29, 2549-2557 (2001).

56. Andrew A. Hill, E.L.B., Maryann Z. Whitley, Greg Tucker-Kellogg, Craig P. Hunter, and Donna K. Slonim Evaluation of normailization procedures for oligonucleotide array data based on spiked cRNA controls. *Genomebiology* 2 (2001).

57. Ott, L. An Introduction to Statistical Methods and Data Analysis, Edn. 4th. (Wadsworth Inc., Belmont, CA; 1992).

58. Lentner, M., Bishop, T. Experimental design and analysis, Edn. 2. (Valley Book Company, Blacksburg; 1986).

59. Fix, E., Hodges, J. Discriminatory analysis, nonparametric discrimination: consistency properties. *Technical report, Randolf Field, Texas: USAF School of Aviation Medicine* (1951).

60. Shai, R., Shi, T., Kremen, T., Horvath, S., Liau, L., Cloughesy, T., Mischel, P., Nelson, S. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene* 22, 4918-4923 (2003).

61. Beer, D.,Kardia, S.,Huang, C.,Giodano, T.,Levin, A.,Misek, D.,Lin, L.,Chen, G.,Gharib, T.,Thomas, D.,Lizyness, M.,Kuick, R.,Hayasaka, S.,Taylor, J., Lannettoni, M.,Orringer, M.,Hanash, S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Naure Medicine* 8, 816-824 (2002).
62. Nguyen, D., Rocke, D. Multiclass cancer classification via partial least squares with gene expression profiles. *Bioinformatics* 18, 1216-1226 (2002).

63. Guo, Q.M. DNA microarray and cancer. *Curr Opin Oncol* 15, 36-43. (2003).

64. Simone, R. Diagnosis and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer* 89, 1599-1604 (2003).

65. Godard, S.,Getz, G.,Delorenzi, M.,Farmer, P.,Kobayashi, Hiroyuki.,Desbaillets, I.,Nozaki, M.,Discerens. A-C.,Hamou, M-F.,Dietrich, P-Y.,Regli, Luca.,Janzer, R.,Bucher, P.,Stupp, R.,Tribolet, N.,Domany, E.,Hegi, M. Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with

angiogenic activity emerges as a strong predictor of subtypes. *Cancer Research* 63, 6613-6625 (2003).

66. Lee, Y., Lee, C.K. Classification of multicategory support vector machines using gene expression data. *Bioinformatics* 19, 1132-1139 (2003).

67. Kim, S., Dougherty, E.,Barrera, J., Chen, Y.,Bittner, M., Trent, J. Strong feature sets from small samples. *Journal of Computational Biology* 9, 127-146 (2002).

68 McShane, L., Shih, J., Michalowska, A. Statistical issues in the design and analysis of gene expression microarray studies of animal models. *Journa of Mammary Gland Biology and Neoplasia* 8, 359 - 373 (2003).

69. Muller, K., Sebastian, M., Ratsch, G., Tsuda, K., Scholkopj, B. An introduction to kernal-based learning algorithms. *IEEE transactions on neural networks* 12, 181-202 (2001).

70. Dougherty, E. Small sample issues fo microarray-based classification. *Comparative and Functional Genomics* 2, 28-34 (2001).

71. Kim, S., Dougherty, E., Shmulevich, I., Hess, K., Hamilton, S., Trent, J., Fuller, G., Zhang, W., Identification of combination gene sets for glioma classification. *Molecular Cancer Therapeutics* 1, 1229-1236 (2002).

72. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* 99, 6567-6572 (2002).

73. Brown, A.G. Nerve Cells and Nervous Systems, Vol. 1, Edn. 2. (Springer-Verlag, London, Berlin, Heidelberg; 2001).

74. Kandel, E., Schwartz, J., Jessell, T. Principles of Neural Science, Edn. 4. (McGraw-Hill, New York; 2000).

75. Kingsley, R.E. Neuroscience, Edn. 2. (Lippincott Williams and Wilkins, Baltimore; 2000).

76. Collins, V.P. Brain tumors: classification and genes. *Journal of Neurological Neurosurgery and Psychiatry* 75, 2-11 (2004).

77. Kleihues, P., Kiessling, M., Janzer, R. Morphological markers in neuro-oncology. *Current Topics in Pathology* 77, 307-338 (1987).
109

78. Cavenee, W.K., Bigner, D., Newcomb, E. in Pathology and genetics: tumors of the nervous system. (eds. P. Klleihues & W.K. Cavenee) 2-9 (International Agency for Cancer Research, Lyon, France; 1997).

79. Davis, R., Kleihues, P., Burger, P. in Pathology and genetics: tumors of the nervous system. (eds. P. Kleihues & W.K. Cavenee) 14-15 (International Agency for Cancer Research, Lyon, France; 1997).

80. Sehgal, A. Molecular Changes During the Genesis of Human Gliomas. *Seminars in Surgical Oncology* 14, 3-12 (1998).

81. Leenstra, S., Oskam, N., Buleveld, E., Bosch, D., Troost, D., Hulesbos, T. Genetic sub-types of human malignant astrocytoma correlate with survival. *International Journal of cancer* 79, 159-165 (1998).

82. Karak, A., Singh, R., Tandon, P., Sarkar, C. A Comparative Survival Evaluation and Assessment of Interclassification Concordance in Adult Supratentorial Astrocytic Tumors. *Pathology Oncology Research* 6, 46-51 (2000).

83. Kleihues, P., Cavenee, W.K. Pathology and Genetics of Tumours of the Nervous System, 2nd ed. (Oxford University Press, New York:; 2000).

84. Brustle, O., Ohgaki, H., Schmitt, H. Primitive Neuroectodermal tumors after prophylactic central nervous system irradiation in children, association with an activated K-ras gene. *Cancer* 69, 2385-2392 (1992).

85. Greenberg, M. Handbook of Neurosurgery. (Greenberg Graphics Inc., Lakeland, Fla; 1997).

86. Bloom, van den J., Wolter, M.,Kuick, R.,Misek, D., Youkilis, A., Wechsler, D., Sommer, C., Reifenberger, G., Hanash, S.Characterization of gene expression profiles associated with glioma progresion using oligonucleotide-based microarray analysis and real-time reverse transcription-polymerase chain reaction. *American Journal of Pathology,* 163, 1033-1043 (2003).

87. Yuen, T., Wurmbach, E., Pfeffer, R., Ebersole, B., Sealfon, S. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleaic Acids Research* 30, e48 (2002).

88. Shi, L Tong, W Fang, H Scherf, U Han, J Puri, R Frueh, F Goodsaid, F Guo, L Su, Z Han, T Fuscoe, J Xu, Z Patterson, T Hong, H Xie, Q Perkins, R Chen, J Dasicano, D. Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* 6(Suppl 2) (2005).

89. Kuo, W., Jenssen, T., Ohno-Machado, O., Kohane, I. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 18, 405-412 (2002).

90. Getz, G., Levine, E., Domany, E. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences* 97, 12079-12084 (2000).

91. Wang, H Malek, R Kwitek, A Greene, A Luu, T., Behdahani, B Fank, B Quackenbush, J Lee, N. Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biology* 4, R5 (2003).

92. Relogio, A., Richter, C.S.A., Ansorge, W., Valcarcel, J. Optimization of oligonucleotide-based DNA microarrays. *Nucleaic Acids Research* 30, e51 (2002).

93. Shalon, D., Smith, S.J., Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hyridization. *Genome Research* 6, 639-645 (1996).

94. Levin A., S.G., and Gutin P. Neoplasms of the Central Nervious System. *Cancer: Principles and Practice of Oncology 3rd ed.*, 1557-1611 (1989).

95. Coons, S., Johnson, P., Scheithauer, B., Yates, A., Perl, D. Improving diagnostic accuracy and interobserver concordance in classification and grading of primary gliomas. *Cancer* 79, 1381-1393 (1997).

96. Blohm, D.H., Guiseppi-Elie, A. New developments in microarray technology. *Current Opinion in Biotechnology* 12, 41-47 (2001).

97. Bretz, F., Landgrebe, J., Brunner, E. Effecient design and analysis of two-color factorial microarray experiments. *Biostatistics* 1, 1-20 (2003).

98. Zar, J. Biostatistical analysis, Edn. 4. (Prentice-Hall, New Jersey; 1999).

99. Cox, W.G., Beaudet, M.P., Agnew, J.Y., Ruth, J.L. Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. *Analytical Biochemistry* 331, 243-254 (2004).

100. Cox, D.R., Oakes, D. Analysis of Survival Data. (Chapman and Hall, London; 1984).

101. Cox, D.R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* 34, 187-220 (1972).

103. Team, R.D.C. (R Foundation for statistical computing, 2003).

104. Liang, Y Diehn, M Watson, N Bollen, A Aldape, K Nicholas, K Lamborn, K Berger, M Botstein, D Brown, P., Isreal, M. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proceedings of the National Academy of Sciences* 102, 5814-5819 (2005).

105. Chiappetta, G., Pentimalli, F., Monaco, M., Fedele, M Pasquinelli, R., Pierantoni, G.M., Ribecco, M., Santelli, G., Califano, D., Pezzullo, L., Fusco, A.. Thymosin beta-10 gene expression as a possible tool in diagnosis of thyroid neoplasias. *Oncology Reports* 12, 239-243 (2004).

106. Santelli, G Califano, D Chiappetta, G Vento, M.T Bartoli, P.C., Zullo, F Trapasso, F Viglietto, G., Thymosin beta-10 gene overexpression is a general event in human carcinogenesis. *American Journal Pathology* 155, 799-804 (1999).

107. Hall, A.K., Hempstead, J., Morgan, J.I. Thymosin beta 10 levels in developing human brain and its regulation by retinoic acid in the HTB-10 neuroblastoma. *Molecular Brain Research* 8, 129-135 (1990).

108. Lee, S.H Son, M.J Oh, S.H Rho, S.B Park, K Kim, Y.J Park, M.S., Lee, J.H. Thymosin (beta) (10) inhibits angiogenesis and tumor growth by interfering with Ras function. *Cancer Research* 65, 137-148 (2005).

109. Zeng, Z.J., Yang, L.Y., Ding, X., Wang, W. Expressions of cysteine-rich61, connective tissue growth factor and Nov genes in hepatocellular carcinoma and their clinical significance. *World Journal of Gastroentrology* 10, 3414-3418 (2004).

110. Wunder, J.S., Eppert, K Burrow, S.R Gokgoz, N Bell, R.S., Andrulis, I.L.. Co-amplification and overexpression of CDK4, SAS and MDM2 occurs frequently in human parosteal osteosarcomas. *Oncogene* 18, 783-788 (1999).

111. Dettling, M. BagBoosting for tumor classification with gene expression data. *Bioinformatics* 20, 3583-3593 (2004).

APPENDIX 1
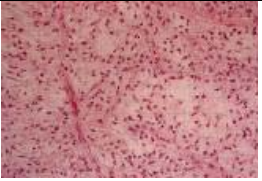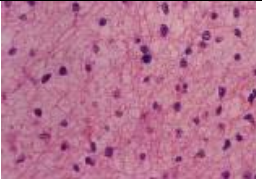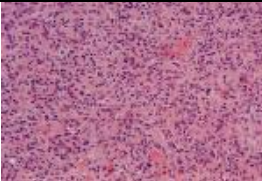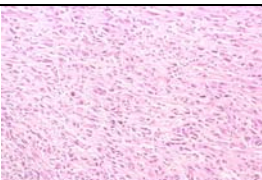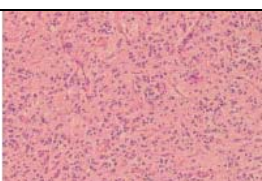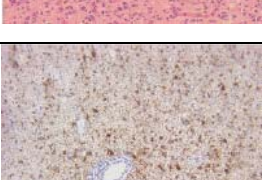
WHO Malignancy Grade	Histological Classification	Histology
Grade I	Pilocystic astrocytoma,	
Grade II	Low grade astrocytoma	
Grade III	Anaplastic astrocytoma	
Grade IV	Glioblastoma Multiforme	
Grade II	Oligodendroglioma	
Grade III	Anaplastic oligodendroglioma	

Table A1. Astrocytoma grades and histological examples of each grade.

APPENDIX 2

Table A2. Control genes added to the MWG 10k A Pan human oligo set

Accession Number	Control Genes
M33197	GAPDH 5'
M33197	GAPDH 3'
X00351	Beta Actin 5'
X00351	Beta Actin 3'
M97935	ISGF 5'
M97935	ISGF 3'
J04422	Bio B
J04423	Bio C
J04424	BioD
X04603	ThrC
M24537	Phe B
M64784	phosphofructokinase, platelet
M27396	asparagine synthetase
M11560	aldolase A, fructose-bisphosphate
XM_083842	phosphoglycerate mutase 1 (brain)
M12996	glucose-6-phosphate dehydrogenase
AB061838	ribosomal protein S3
XM_088688	non-POU-domain-containing, octamer-binding
NM_002954	ribosomal protein S27a
NM_005566	lactate dehydrogenase A
NM_000291	phosphoglycerate kinase 1 G
NM_014763	mitochondrial ribosomal protein L19
AA453756	Rho GDP dissociation inhibitor (GDI) alpha
NM_005566	lactate dehydrogenase A
NM_004048	beta-2-microglobulin
M64784	phosphofructokinase, platelet
M11560	aldolase A, fructose-bisphosphate
NM_002954	ribosomal protein S27a

.

APPENDIX 3

Table A3. Brain tumor related genes added to the 10KO Table

1	NM_004341_1	carbamoylphosphate synthetase 2/aspartate transcarbamylase/dihydroorotase
2	X86098_1	BS69 protein
3	K01396_1	H alpha-1-antitrypsin mRNA. 11/1994"
4	NM_022111_1	claspin; CLSPN
5	X95735_1	zyxin.
6	BC010577_1	granulin, mRNA (cDNA clone MGC:9342 IMAGE:3457813)
7	U67963_1	lysophospholipase homolog (HU-K5)
8	J05243_1	nonerythroid alpha-spectrin (SPTAN1)
9	J02611_1	apolipoprotein D mRNA
10	J04599_1	hPGI mRNA encoding bone small proteoglycan I (biglycan)
11	M17783_1	glia-derived nexin (GDN) mRNA, 5' end. 4/1993"
12	U20498_1	p19 protein mRNA, complete cds. 1/1996"
13	NM_001878_1	cellular retinoic acid binding protein 2; CRABP2
14	M10905_1	cellular fibronectin mRNA. 11/1994
15	L07493_1	replication protein A 14kDa subunit (RPA)
16	NM_020349_1	ankyrin repeat domain 2; ANKRD2
17	NM_018659_1	cytokine-like 1; CYTL1
18	M34458_1	lamin B mRNA
19	M87339_1	replication factor C, 37-kDa subunit mRNA, complete cds. 10/1996"
20	X77584_1	ATL-derived factor/thiredoxin. 9/2004
21	M28215_1	GTP-binding protein (RAB5) mRNA, complete cds. 1/1995
22	D00632_1	glutathione peroxidase, complete cds. 3/1998
23	Z21507_1	EF-1delta gene encoding human elongation factor-1-delta. 1/1994
24	NM_052951_1	terminal deoxynucleotidyltransferase interacting factor 1; DNTTIP1
25	X74794_1	P1-Cdc21 mRNA. 9/1996
26	NM_001810_1	centromere protein B; CENPB
27	J04164_1	interferon-inducible protein 9-27 mRNA, complete cds. 4/1993"
28	X77588_1	TE2 mRNA for ARD-1 N-acetyltransferase homologue. 7/1994
29	X62534_1	HMG-2 mRNA. 11/1991
30	M61764_1	gamma-tubulin mRNA, complete cds. 11/1994
31	U51477_1	diacylglycerol kinase zeta mRNA, complete cds. 5/1996
32	M23178_1	homologue-1 of gene encoding alpha subunit of murine cytokine (MIP1/SCI),
33	NM_016382_1	CD244 natural killer cell receptor 2B4; CD244
34	L29222_1	clk1
35	U90551_1	histone 2A-like protein (H2A/l)

Table A3. continued. Brain tumor related genes added to the 10KO

	Accession #	Gene Name
36	NM_032493_1	adaptor-related protein complex 1, mu 1 subunit; AP1M1
37	U52828_1	delta-catenin mRNA,
38	U33632_1	two P-domain K+ channel TWIK-1
39	X53793_1	ADE2H1 mRNA showing homologies to SAICAR synthetase
40	X60486_1	H4/g gene for H4 histone.
41	NM_032545_1	cryptic; CFC1
42	NM_033259_1	CaM-KII inhibitory protein; CAM-KIIN
43	U52100_1	XMP
44	NM_004244_1	CD163 antigen isoform a; CD163
45	NM_001826_1	CDC28 protein kinase 1B; CKS1B
46	X51405_1	for carboxypeptidase E (EC 3.4.17.10). 3/1995
47	NM_001645_1	apolipoprotein C-I precursor; APOC1
48	AF010314_1	Pig10 (PIG10) mRNA, complete cds. 1/1998"
49	X97324_1	adipophilin. 3/1997
50	NM_138455_1	collagen triple helix repeat containing 1; CTHRC1
51	NM_001323_1	cystatin M precursor; CST6
52	U73379_1	cyclin-selective ubiquitin carrier protein mRNA
53	X69838_1	G9a.
54	X82434_1	emerin.
55	M36711_1	sequence-specific DNA-binding protein (AP-2) mRNA,
56	L14542_1	lectin-like type II integral membrane protein (NKG2-E) mRNA,
57	NM_052842_1	BCL2-like 12 isoform 2; BCL2L12
58	U07358_1	protein kinase (zpk) mRNA, complete cds. 5/1995
59	U21090_1	DNA polymerase delta small subunit mRNA, complete cds. 10/1995
60	D89667_1	mRNA for c-myc binding protein, complete cds. 2/1999"
61	U33267_1	glycine receptor beta subunit (GLRB) mRNA, complete cds. 12/1996
62	NM_130468_1	dermatan 4 sulfotransferase 1; D4ST1
63	X54942_1	ckshs2 mRNA for Cks1 protein homologue. 4/1992
64	NM_005209_1	crystallin, beta A2; CRYBA2"
65	M91670_1	ubiquitin carrier protein (E2-EPF) mRNA, complete cds. 12/1994
66	M55542_1	guanylate binding protein isoform I (GBP-2) mRNA, complete cds. 4/1993
67	M94345_1	macrophage capping protein mRNA, complete cds. 1/1995"
68	U46744_1	dystrobrevin-alpha mRNA, complete cds. 4/1996
69	U65932_1	extracellular matrix protein 1 (ECM1) mRNA, complete cds. 8/1996
70	U79299_1	neuronal olfactomedin-related ER localized protein mRNA
71	U28386_1	nuclear localization sequence receptor hSRP1alpha
72	NM_052999_1	chemokine-like factor superfamily 1 isoform 13; CKLFSF1
73	X16841_1	for a nontransmembrane isoform of N-CAM from skeletal muscle. 9/2004
74	X78565_1	for tenascin-C, 7560bp. 5/1995"
75	X66360_1	PCTAIRE-2 for serine/threonine protein kinase. 1/1993
76	L37936_1	nuclear-encoded mitochondrial elongation factor Ts (EF-Ts)
77	NM_080603_1	zinc finger, SWIM domain containing 1; ZSWIM1
78	U43885_1	Grb2-associated binder-1

APPENDIX 4

Table A4. NQC data for arrays hybridized in the brain tumor study.

Array	Lack of Fit p-val	3'5'ratio (cutoff: 0 - 3)		Array	Lack of Fit p-val	3'5'ratio (cutoff: 0 - 3)	
	a=.01	GAPDH	Beta Actin		a=.01	GAPDH	Beta Actin
aa1_12-7	0.9995219	1.715162	1.456302	gm24_12-14	0.01808889	1.456864	1.559403
aa2_12-32	0.2656423	2.978236	1.657802	gm27_12-13	0.7752857	1.598357	1.368301
aa2_13-36	0.651442	1.669108	1.032857	gm27_12-49	0.7982792	2.639673	1.151611
aa3_13-46	0.548669	1.900667	1.254248	gm27_13-37	0.9814702	2.131558	0.9429497
aa5_13-20	0.9598248	1.226053	0.9680632	gm28_12-35	0.9324798	1.2982	1.003555
aa6_12-19	0.0733892	1.650434	2.480799	gm29_13-21	0.999739	1.739052	1.224106
aa8_13-14	1	1.647383	0.9458296	gm30_12-9	0.4023522	1.851453	1.100206
aa9_12-40	0.1151365	1.46014	1.445064	gm39_13-15	0.6430087	1.473216	1.637815
aa10_13-24	0.5656413	3.719946	1.0987	gm40_13-9	0.3262023	1.804772	5.267451
aa11_12-33	0.9999824	1.360714	0.9569297	gm47_13-16	0.995306	2.031892	1.484073
aa11_13-13	0.8125632	2.768844	2.109789	gm49_13-17	0.6745692	1.70496	3.314934
ao1_12-45	0.5107544	11.92174	2.498547	gm50_13-31	0.3633336	1.656788	4.433658
ao2_12-23	0.79086	3.50607	1.128656	ol1_13-45	0.983258	1.841686	1.381023
ao3_12-39	0.9462462	4.009255	1.0047	ol2_12-6	0.01362917	2.522023	2.217825
ao4_13-39	0.9865006	1.769952	1.608386	ol3_12-3	0.2206988	3.522815	1.009913
ao5_13-22	0.9996987	2.981656	1.387635	ol4_12-42	0.9677053	4.440736	1.892986
ao7_12-34	0.04068426	1.713749	1.176457	ol5_13-29	0.9992544	1.567363	2.029612
ao7_13-41	0.9616176	1.352545	1.293315	ol6_13-25	0.8150847	4.874391	1.07875
ao8_12-11	0.9394637	2.465128	1.014478	ol7_12-16	0.946334	1.408027	2.756842
ao9_12-10	0.999991	1.120634	0.9559904	ol7_13-30	0.17698	2.071088	3.364591
ao12_12-2	0.998246	1.620303	1.214146	ol9_12-41	0.9424667	4.009447	10.95967
ao13_12-46	0.9618674	1.064423	1.004983	ol9_13-19	0.5520059	1.583125	2.462907
gm1_13-6	0.2667973	4.304432	1.560787	ol10_13-8	0.9989395	15.16516	15.68594
gm2_13-7	0.767172	3.662575	1.419076	ol13_12-44	0.9999803	4.015749	3.029822
gm3_13-2	0.988722	5.034602	1.458833	nb1_13-48	1	1.598918	0.8760355
gm4_13-4	0.9977209	1.250141	1.183477	nb2_96r_12-26	1	1.153371	1.241361
gm6_12-17	0.9221156	2.781463	1.19103	nb2_206d_13-4	0.7472291	1.891704	2.187962
gm6_12-25	0.2054556	5.908584	1.30218	nb5_12-8	0.9998686	0.9762596	1.256204
gm6_13-43	1	1.442977	1.255034	pa1_13-4	0.989558	1.436791	0.9304682
gm7_12-27	0.9451391	2.480818	1.487407	pa2_12-31	0.5321859	7.78268	3.100132
gm8_13-28	0.9877044	2.408643	1.622863	pa3_12-20	0.1540523	2.696343	1.036984
gm9_13-15	0.244511	2.268389	1.081125	pa4_12-20	0.9686794	4.695413	1.107287
gm10_12-37	0.9995305	4.665188	5.821688	pa5_13-10	0.6553276	4.445429	2.793407
gm11_12-5	0.001310677	1.152852	1.633825	pa7_12-21	0.951083	3.215228	1.588234
gm11_12-23	0.783075	3.44635	1.099151	pa8_13-33	0.3769831	2.09212	1.074204
gm13_12-29	0.999378	0.9791048	15.53818	pa9_13-35	0.4725177	1.948792	1.1817
gm15_12-18	0.9451391	2.480818	1.487407	pa10_12-30	0.7617267	2.285632	2.665949
gm17_13-32	0.1621701	2.757798	1.97126	pa10_12-43	0.08744103	7.607161	1.935607
gm18_12-35	0.6472216	1.78984	1.087688	pa10_13-36	0.999757	5.629308	1.474801
gm19_12-22	0.8847309	2.110881	1.081197	pa11_12-15	0.2739063	1.965574	1.060384

APPENDIX 5

Table A5. Sample distrubution and survival times for brain tumors assayed.

Index	Histology	Days	Censor Status	Index	Histology	Days	Censor Status
1	aa1	69	d	35	pa1	2555	c
2	aa2	583	d	36	pa2	1599	c
3	aa3	735	c	37	pa3	744	c
4	aa5	2151	d	38	pa4	2249	c
5	aa6	1526	c	39	pa5	2402	c
6	aa8	69	d	40	pa7	2362	c
7	aa9	553	d	41	pa8	1872	c
8	aa10	84	d	42	pa9	1851	c
9	aa11	913	c	43	pa10	1008	c
10	gm1	741	d	44	pa11	381	c
11	gm2	462	d	45	ol1	2775	c
12	gm3	469	d	46	ol2	2084	d
13	gm4	567	d	47	ol3	1968	d
14	gm6	98	d	48	ol4	2293	c
15	gm7	2061	c	49	ol5	2229	c
16	gm8	78	d	50	ol6	2187	c
17	gm9	300	d	51	ol7	1870	c
18	gm10	336	d	52	ol9	1483	d
19	gm11	70	d	53	ol10	1511	d
20	gm13	25	d	54	ol13	300	d
21	gm15	1897	c	55	ao1	32	d
22	gm17	1103	c	56	ao2	2297	c
23	gm18	131	d	57	ao3	1739	c
24	gm19	335	d	58	ao4	1732	c
25	gm24	467	d	59	ao5	?	
26	gm27	132	d	60	ao7	609	d
27	gm28	107	d	61	ao8	177	d
28	gm29	430	d	62	ao9	15	d
29	gm30	305	d	63	ao12	83	d
30	gm39	421	d	64	ao13	?	
31	gm40	180	d				
32	gm47	189	d				
33	gm49	335	d				
34	gm50	274	d				

APPENDIX 6
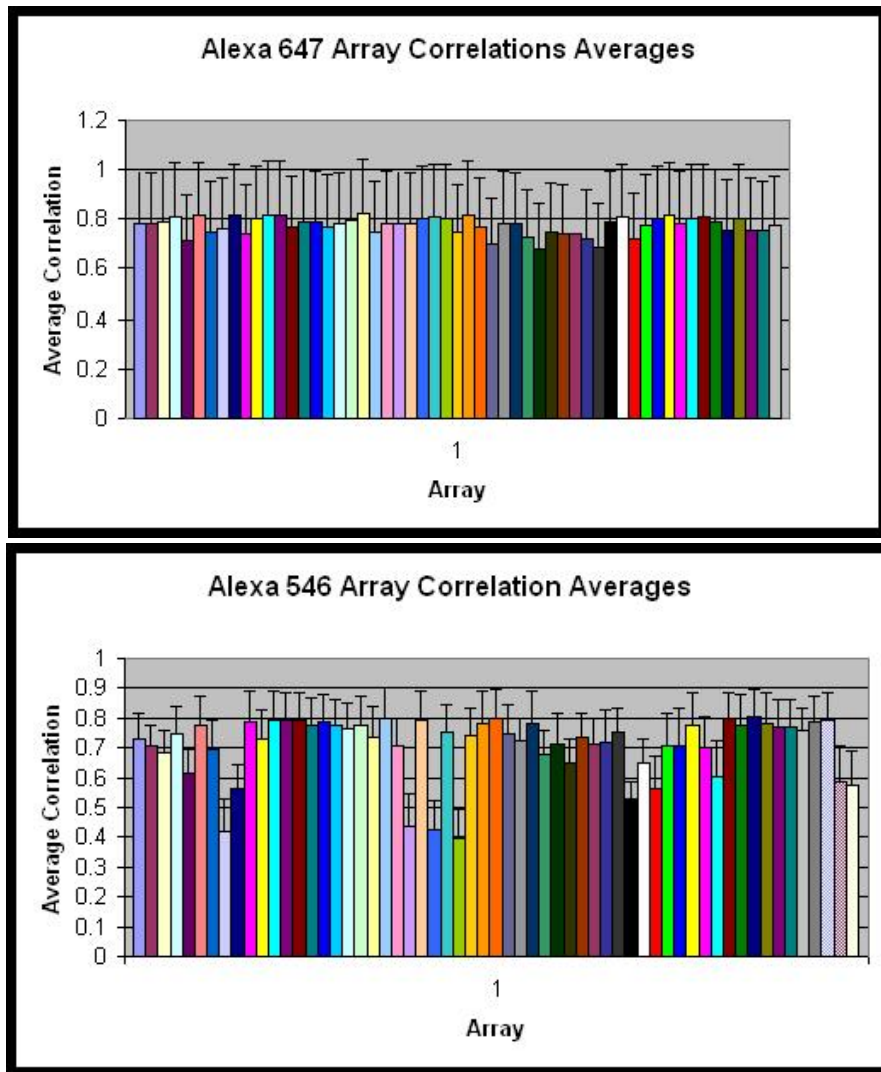
Figure A6. Correlation averages for Alexafluor 647 (top) and Alexafluor 546 (bottom) channels for C3B arrays. Alexafluor 647 channel was the reference channel, and the Alexafluor 546 channel was the sample channel. The x-axis represents individual arrays and the y-axis represents calculated correlation averages for all pair-wise array correlations. Error bars represent the standard deviation among the pair-wise correlation calculations.

APPENDIX 7

Table A7 Significant gene related to survival.

[A]Gene Panel	Gene	d0 Exptected	d0 Observed
	NM_005981_1 sarcoma amplified sequence SAS	-0.023879	-3.242619
	NM_004172_1 solute carrier family 1 glial high affinity glutamate transporter	0.8284317	-3.022238
	NM_014699_1 KIAA0296 gene product KIAA0296	0.1773652	-2.924185
	XM_071619_1 similar to Ser Arg related nuclear matrix protein plenty of prolines 101	0.5690327	-2.75241
	NM_000582_1 secreted phosphoprotein 1	0.0380829	-2.734897
	NM_004059_1 cysteine conjugate beta lyase	0.0224218	-2.656671
	NM_001885_1 crystallin alpha B CRYAB	0.2697162	-2.490686
	NM_022082_1 hypothetical protein FLJ23412 FLJ23412	-0.642498	-2.297341
	NM_018299_1 hypothetical protein FLJ11011 FLJ11011	0.3351757	-2.27485
	NM_002489_1 similar to NADH dehydrogenase ubiquinone 1 alpha subcomplex	-0.516387	-2.2423
	NM_000773_1 cytochrome P450 subfamily IIE ethanol inducible CYP2E	0.0147565	-2.225285
	XM_038988_1 basement membrane induced gene ICB 1	0.6088045	-2.221171
	XM_011281_1 hypothetical protein PRO1580 PRO1580	0.0369406	-2.217762
	NM_018340_1 similar to hypothetical protein FLJ11151	-0.523155	-2.214486
	NM_021126_1 mercaptopyruvate sulfurtransferase MPST	-0.581941	-2.210816
	NM_052842_1 BCL2-like 12 isoform 2; BCL2L12 1	1.8954514	-2.202491
	BC014644_1 similar to high mobility group nonhistone chromosomal protein 17	-1.153229	-2.196156
	NM_000272_1 nephronophthisis 1 juvenile NPHP1	0.4966613	-2.176307
	NM_002433_1 myelin oligodendrocyte glycoprotein MOG	-0.241989	-2.156389
	NM_053036_1 neuropeptide G protein coupled receptor neuropeptide	0.0450747	-2.154447
	NM_003884_1 p300 CBP associated factor PCAF	0.0466943	-2.152312
	NM_001586_1 chromosome X open reading frame 2 CXORF2	0.0877754	-2.134287
	NM_016610_1 Toll like receptor 8 LOC51311	-0.644839	-2.13223
	NM_020162_1 hypothetical protein DKFZp762F2011 DKFZp762F2011	0.1585426	-2.124117
	NM_005166_1 amyloid beta A4 precursor like protein 1 APLP1	-0.145051	-2.108784

APPENDIX 7 CONT.

NM_006574_1 chondroitin sulfate proteoglycan 5 neuroglycan C CSPG5	0.4493752	-2.10041
XM_017093_1 hypothetical protein FLJ23537 FLJ23537	1.2634244	-2.099067
NM_018351_1 hypothetical protein FLJ11183 FLJ11183	0.0017572	-2.066803
NM_016941_1 delta Drosophila like 3 DLL3	0.1262889	-2.063305
NM_004765_1 B cell CLL lymphoma 7C BCL7C	0.1735181	-2.055361
NM_001446_1 similar to fatty acid binding protein 7 brain	0.3582606	-2.049895
NM_006812_1 amplified in osteosarcoma OS 9	-0.020539	-2.048796
NM_000788_1 deoxycytidine kinase DCK	0.430469	-2.046858
NM_002519_1 similar to nuclear protein ataxia telangiectasia locus	-0.269328	-2.045158
NM_022783_1 hypothetical protein FLJ12428 FLJ12428	-0.341256	-2.043382
NM_001828_1 Charot Leyden crystal protein CLC	-0.560267	-2.042737
XM_007370_1 estrogen receptor 2 ER beta ESR2	-0.045282	-2.031634
NM_004808_1 N myristoyltransferase 2 NMT2	0.3063358	-2.022434
NM_019052_1 HCR a helix coiled coil rod homologue HCR	0.7611428	-2.002306
NM_003671_1 CDC14 cell division cycle 14 S cerevisiae homolog B	-0.242518	-2.002179
NM_016282_1 adenylate kinase 3 alpha like AKL3L	0.6670356	-1.978206
NM_004717_1 diacylglycerol kinase iota DGKI	0.0243335	-1.977256
XM_071577_1 acid fibroblast growth factor like protein GLIO703	1.488031	-1.961508
NM_025150_1 hypothetical protein FLJ12528 FLJ12528	-0.387669	-1.955981
NM_001683_1 ATPase Ca++ transporting plasma membrane 2 ATP2B2	-0.227049	-1.948935
NM_000533_1 proteolipid protein P	0.0752973	-1.936391
NM_013364_1 paraneoplastic cancer testis brain antigen MA5	0.085064	-1.930121
NM_024700_1 hypothetical protein FLJ12553 FLJ12553	-0.360248	-1.92869
NM_002612_1 pyruvate dehydrogenase kinase isoenzyme 4 PDK4	0.3506014	-1.922289
NM_016376_1 ANKHZN protein ANKHZN	-0.63335	-1.92225
XM_009122_1 selenoprotein W 1 SEPW1	1.5054079	-1.92169
NM_018962_1 Down syndrome critical region gene 6 DSCR6	0.0990775	-1.916915
NM_004877_1 glia maturation factor gamma GMFG	0.123034	-1.90938
NM_002734_1 protein kinase cAMP dependent regulatory type I	-0.096815	-1.904077
NM_017821_1 hypothetical protein FLJ20435 FLJ20435	0.070982	-1.902981
NM_006454_1 Mad4 homolog MAD4	0.0414242	-1.888498
NM_002928_1 regulator of G protein signalling 16 RGS16	0.5035423	-1.87524
NM_024654_1 hypothetical protein FLJ23323 FLJ23323	-0.351743	-1.873864
NM_016185_1 hematological and neurological expressed 1 HN1	-0.428936	-1.869245

APPENDIX 7 CONT.

	NM_006359_1 solute carrier family 9 sodium hydrogen exchanger isoform 6	-0.202375	-1.862702
	NM_004738_1 VAMP vesicle associated membrane protein associated protein B and C	-0.143195	-1.859751
	NM_000898_1 monoamine oxidase B MAOB	-0.183845	-1.848203
	NM_006991_1 similar to ADP ribosylation factor like 4	-0.462676	-1.845542
	NM_006770_1 macrophage receptor with collagenous structure MARCO	0.0468516	-1.843902
	NM_016415_1 clone FLB3816 LOC51216	-0.898242	-1.814268
	XM_071571_1 similar to clone FLB3816 H sapiens LOC82147	0.56164	-1.81148
	NM_021927_1 hypothetical protein FLJ13220 FLJ13220	-0.224213	-1.807185
	NM_024330_1 hypothetical protein MGC4365 MGC4365	-0.428338	-1.805333
	NM_021784_1 hepatocyte nuclear factor 3 beta HNF3B	-0.654358	-1.802306
	NM_000083_1 chloride channel 1 skeletal muscle	0.3619369	-1.801662
	NM_017601_1 similar to hypothetical protein DKFZp761H221	-0.444022	-1.797454
	NM_000744_1 cholinergic receptor nicotinic alpha polypeptide 4 CHRNA4	-0.14302	-1.797368
	NM_001648_1 kallikrein 3 prostate specific antigen KLK3	-1.001829	-1.793426
	NM_005178_1 B cell CLL lymphoma 3 BCL3	-0.574258	-1.79341
	NM_004268_1 cofactor required for Sp1 transcriptional activation	0.6405076	-1.786783
	NM_000090_1 collagen type III alpha 1 E	0.494296	-1.783971
	NM_001672_1 agouti mouse signaling protein ASIP	0.1130363	-1.783246
	NM_022552_1 DNA cytosine 5 methyltransferase 3 alpha DNMT3A	-0.383603	-1.779418
	NM_004666_1 vanin 1 VNN1	0.3738589	-1.772467
	BC009026_1 ligase III DNA ATP dependent LIG3	-0.599535	-1.771561
	NM_001160_1 apoptotic protease activating factor APAF1	-0.573522	-1.765291
	NM_025105_1 hypothetical protein FLJ12409 FLJ12409	-0.339748	-1.762865
	NM_006716_1 hypothetical gene supported by NM_006716 LOC82512	-0.527153	-1.762787
	NM_001902_1 similar to cystathionase cystathionine gamma lyase	0.5262126	-1.75797
	NM_021098_1 similar to calcium channel voltage dependent alpha 1H subunit	-0.423837	2.079795
	NM_006293_1 TYRO3 protein tyrosine kinase TYRO3	0.2080335	2.0850642
	NM_004270_1 cofactor required for Sp1 transcriptional activation subunit 9 33kD	0.7962818	2.1021187
	NM_000342_1 solute carrier family 4 anion exchanger member 1	0.1673858	2.1021528
	NM_005688_1 ATP binding cassette sub family C CFTR MRP member 5	0.9328213	2.1118648
	NM_025134_1 hypothetical protein FLJ12178 FLJ12178	-0.31455	2.119763
	NM_000852_1 glutathione S transferase pi GSTP1	0.2686803	2.1588387
	NM_014588_1 visual system homeobox 1 zebrafish homolog CHX10 like	-0.296528	2.1772947
	NM_018496_1 similar to hypothetical protein PRO0889	-1.230039	2.218164

APPENDIX 7 CONT.

🟥	NM_018358_1 hypothetical protein	0.9563472	2.2672145
🟥	NM_022473_1 zinc finger protein 106 ZFP106	-0.747348	2.2685475
🟥	NM_005500_1 SUMO 1 activating enzyme subunit 1 SAE1	-0.580502	2.3193085
🟥	NM_004341_1 carbamoylphosphate synthetase 2	0.6151074	2.359525
🟥	NM_025098_1 hypothetical protein FLJ22644 FLJ22644	-0.3233	2.3829834
🟩	NM_001554_1 cysteine rich angiogenic inducer 61 CYR61	0.5356279	2.452571
🟩	NM_017818_1 hypothetical protein FLJ20430 FLJ20430	0.5403622	2.5232244
🟩	NM_024663_1 hypothetical protein FLJ11583 FLJ11583	-0.323477	2.6286561
🟩	NM_006636_1 methylene tetrahydrofolate dehydrogenase NAD+ dependent	0.5007502	2.6340063
🟩	NM_004537_1 nucleosome assembly protein 1 like 1 NAP1L1	0.2572645	2.7493677

[A]Green rows were genes that made up the 10 gene panel, red the 20 gene panel, blue the 50 gene panel, and pink the 100 gene panel such that each increasing gene panel contained genes from the reduced panels, for example, the 20 gene panel contained 10 green gene rows and 10 red gene rows. The white rows were excluded from the analysis. Cells highlighted in grey are genes that were common betweent the C3B and SMD arrays.