An Adenocarcinoma Case Study of the BaFL Protocol: Biological Probe Filtering for Robust Microarray Analysis

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Kevin Thompson
Masters of Science
University of Wisconsin - Milwaukee, 2002

Director: Dr. Jennifer Weller, Assoc. Professor
College of Computing and Informatics

Fall Semester 2008
George Mason University
Fairfax, VA

DEDICATION

The work presented in this dissertation is dedicated to my parents, Virgil and Rita Thompson, who I can never repay all of their support.

ACKNOWLEDGEMENTS

Dr. Jennifer Weller my committee advisor and mentor, your guidance was always appreciated. Your persistence was needed and hopefully the work measures to your expectations.

My committee members: Dr. Kinser for a wonderful open door policy, which I abused. Dr. Solka thank you for all of your advice and the interest you've taken in my dissertation, I hope that you are proud of the work. Dr. Willett for his support and the constant reminders about the non-linearity of the system.

My wonderful friends that I have met at this institution, their companionship will never be forgotten. The additional faculty and staff within our department: Dr. Saleet Jafri, Glenda Wilson, Chris Ryan, and Mary Margaret Flannery.

Qui est per omnia secula benedictus: Larry Thompson, Merrit Nichols, Stuart Daily, Brian Bowers, Kim Pederson, Simon Bailey

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

AN ADENOCARCINOMA CASE STUDY OF THE BAFL PROTOCOL:
BIOLOGICAL PROBE FILTERING FOR ROBUST MICROARRAY ANALYSIS

Kevin Thompson, PhD

George Mason University, 2008

Dissertation Director: Dr. Jennifer Weller

Microarrays are high throughput data measurement technologies; those that assay gene expression levels allowing investigators to simultaneously estimate the level of thousands of cellular transcripts present in a sample at the time of collection. Many sources of variation have plagued Microarray analysis, leading to apparent inconsistencies between experimental results derived from independent platforms. A rigorous, robust set of methods for identifying all of the currently known sources of variability and consistently applying them across large data sets has been implemented in the Biologically applied Filter Level, BaFL, protocol. This protocol eliminates all probes for which the underlying sequence characteristics are missing, because of which the probe characteristics, including the identification of the measured transcript region, are impossible to derive. The remaining probes are processed through the biophysical software to determine their Gibb's free energy, as a measure of the solution stability. This measure eliminates any overly stable probes, which would be less assessable to measure the desired transcript region. The filtering process also enforces a range of acceptable signal intensity measurements, the result of scanner characteristics. Measurements outside the linear range

violate the linear correlation relationship between transcript concentration and signal intensity. Probes identified as covering single nucleotide polymorphisms are identified and removed. The Ensembl database is queried to identify probes which measure single specific gene transcript regions, all other probes were excluded. The final step is to enforce a rule that a minimum of four probes are retained, so that any given statistical estimator of concentration has an adequate basis. Samples are subject to many technical steps, so tests for outliers are implemented that included comparisons of representative probe intensities and probe numbers, against the population mean. Samples exceeding ±2 standard deviations of the average probe numbers and probe intensities are removed. ProbeSet constituents at this stage may not be identical across all samples, with differences arising from the linear range filter step. By performing an intersection operation of the remaining probes across all samples, still enforcing a minimum of four probes per ProbeSet, a final, common ProbeSet dataset is derived, which is used as the basis of all further comparisons and analyses.

The suggested data models demonstrated improved performance across three classification algorithms, and remarkable latent structure can be seen across the data models. When Bonferonni correction is applied and the intersecting genes identified a final candidate gene list of 30 ProbeSets results. By including on/off genes in the list, an additional ProbeSet is identified. These 31 candidate genes demonstrate notable connectivity in their GO and KEGG associations. Literature review of the genes establishes that these associations arise from properties specific to angiogenesis and tumorogenesis. A multiclass dataset of non small cell lung cancer samples was constructed and information gain calculated from the k-means clustering efficiency. A candidate list of 18 genes is shown to possess an information gain greater than or equal to 0.8. The literature review of these 18 genes provides evidence that abnormal cytokinesis may underlie

tumorogenesis for both cancer sub-types. The squamous cell carcinomas, in particular, appear to suffering from the production of radical oxidative species.

Currently most Microarray analyses implement one of a small number of published probe cleansing algorithms. Occasional efforts to accommodate one of the confounding factors of the probe-transcript interaction have been made, but no method is as inclusive as that presented in this work. Further, no work exists that demonstrates the improved efficacy of removing a factor on subsequent performance with the existing algorithms. Great effort has been taken here to show that analysis of the resulting datasets leads to greatly improved consistency in inter-experimental comparisons, using two independent lung adenocarcinoma datasets, in comparison to the pre-eminent probe cleansing methodologies, RMA and dCHIP.

Chapter 1: An Introduction to Microarrays

Gene expression Microarray technology is a high throughput capture detection assay that produces a fairly complete representation of a sample's genome-wide transcript complement [1, 2]. The assay has been predominately utilized for gene expression experiments, but modifications in probe design allow the use of this platform for single nucleotide polymorphism and comparative genome hybridization experiments [3-5]. The array concept is a capture detection assay as developed from Northern and Southern blots [5, 6], and this detection method has spawned analogous methods such as the immunological assays known as ELISAs (enzyme linked immunosorbent assays) [7]. Unfortunately, for most platforms, rigorous experimental design has been sacrificed in order to rapidly achieve high throughput genome-wide analysis, but we have developed a number of assessment strategies for the individual probes that allow us to ascertain the reliability of individual measurements.

Our focus here is upon gene expression analysis, and particularly those performed using the Affymetrix platform [8], which has the benefits of sampling transcripts multiple times, and having probes short enough to be sensitive to individual differences in sequence [8, 9]. For gene expression Microarray analysis, mRNA transcripts are extracted from a collection of cells, which can represent anything from a pure culture to a cryogenically preserved and laser-dissected tissue sample. These mRNA samples can be sheared, transformed or otherwise amplified into cDNA or cRNA 'targets', and at some part of the protocol they are labeled, most often with a fluorescent dye [5].

Probes are the complementary strands intended to hybridize to these targets, and are designed against sequence databases of the genes of interest; chemical synthesis may occur on the array surface (as for Affymetrix arrays) or occur first and then be subsequently attached at specific positions [5]. There is no fixed limitation on probe length, and indeed the probes on an array may differ slightly in length, if a melting temperature has been the defining parameter in the design process [10]. Arrays are typically divided into short oligo and long oligo categories, and cDNA arrays. Commercial vendors currently offer oligonucleotide-based arrays, ranging from 24-mers [11] and 25-mers [8] to 70-mers [12], while cDNA arrays (usually, in fact, the PCR product of cDNA inserts) were prevalent in the 1990s and early 2000's [2, 13-15].

The difficulties inherent in interpreting cDNA data have led to a considerable drop in the number of publications that use this style of array but they typically had much longer probes, usually 100+ nucleotides [15]. Each probe design strategy has its own individual merit with respect to a particular experimental goal [14, 16], but the short oligonucleotide platforms are the most flexible [10] and, as costs have decreased, have become the most prevalent in published scientific literature.

Target mixtures are applied to the array surface and, after sufficient hybridization time and removal of non-specifically bound material, the elicitation/capture of the signal molecule occurs [10]. For fluorescent dyes this requires laser excitation of the dye in an imager or scanner with sufficient resolution to separate the spots [5, 14]. From these pixel-by-pixel scanner measurements an overall fluorescence per spot provides a qualitative representation of each gene's relative abundance [14]. A common assumption is that the intensity of the spot correlates to the concentration of the original mRNA, directly and consistently across all spots. However, as

previously stated, rigorous reagent design, in a large number of experiments, has been scarce, especially in comparison to the family of immunological assay technologies.

ELISAs use a similar biomolecule-based capture detection assay; however, ELISAs are made quantitative by the nature and number of controls that are standard practice for the assays. As quantitative assays ELISAs are performed with background controls, a dilution series standard, sample replicates, and in addition, scanner limits are used to set boundaries on the measurements when interpreting the results [7]. The majority of these features is absent or unrecognized in commercial Microarray platforms and their associated analysis pipelines, although each has been the subject of one or more studies by individual investigators and shown to have considerable impact on the outcome [17-20].

The work reported here attempts to find a computational remedy for each of the weak design points, in order to derive a more reliable representation of the transcript concentration. The procedures developed are demonstrated in detail for the Affymetrix platform, but they are extendable to any platform. The next sections present background material about the probe-target reaction, the various measurement platforms, and gene expression experiments. Once the apparent weak points have been identified, the important question is the extent to which they change and whether they improve the interpretation of the data, which we approach in a number of ways described in detail below. The final section lists the specific aims of the research described in this dissertation.

Microarrays: Hybridization and Signal

DNA Microarray assays are based upon the principal of complementary nucleotides binding in solution to form a stable anti-parallel duplex [10, 21]. Hydrogen bonds dictate the kinetics of the nucleotide pairings, and base-stacking interactions govern the stability of the resultant product [22-24]. The factors affecting the rate and extent of these endothermic reactions have been extensively studied, including the ionic conditions, temperature and solvent characteristics under which the single strands will form hybrids [22, 24-27]. However an important and not well understood distinction in a Microarray experiment is that the probes are fixed to a surface and may not freely diffuse [10, 14]; in addition the reactions are not composed of single sets of reactants but a highly multiplex set of unimolecular and bimolecular reactions within and between probes and targets.

That is, since a heterogeneous mixture of transcript targets exists, duplex reactions compete to form stable compounds, some containing a variety of mismatches that dissociate slowly [28], and therefore, to optimize the measurement of the best matches, the mixture of reactions must be allowed to reach a state of equilibrium [3, 10]. The formation of sub-optimal duplexes is termed cross-hybridization, and those targets with high enough sequence similarity to compete for stable products are 'false positive' signals under the usual guidelines of interpretation [17, 20]. The combination of desired and undesired hybridizations generates a fluorescent signal, recorded as a spot of high intensity that is extracted by the instrument's scanning software. The amount and position of dye incorporated into the target will clearly also affect the intensity of the signal and must be accounted for in quantitative interpretations [5].

Factors influencing Signal Interpretation

Hybridization Factors

Microarray experiments result in an image output recording the extent of millions of uncatalyzed chemical reactions; these reactions are influenced by reactant concentrations, solvent conditions, time, and temperature [10]. In addition to these basic reaction parameters the reagents themselves have properties which affect the reaction, as discussed below:

1) Probe Concentration. Probes are affixed to their designated grid positions in high concentrations, in order to drive the reaction towards product formation [14]. This is necessary since the sample's target concentration is unknown and often times low. In addition to concentration, individual probe attributes such as length and base composition and order affect the stability of the product [10]. Probe length limits the concentration of probes which can be affixed since individual probes must not interact with one another and there must be sufficient space for target diffusion between the probe molecules [29]. Longer products, those over ~70 nucleotides in length, approach a common melting temperature and have little sequence dependence, while shorter duplexes demonstrate more variability in temperatures and are influenced by base composition and order as well as length [10, 22, 25, 26]. Self complementary pairings lead to the formation of secondary structure in both the probes and targets, which compete for maximal product duplex formation [30]. Probe secondary structure formation becomes increasingly likely as the probe length increases. The stability of a duplex is more sensitive to mismatches when the duplex is shorter, which can be treated as either a confounding factor or a desirable feature,

depending on the sophistication of the controls and analysis methods [10]. One aspect our group has given particular attention to is the detection of SNPs in the 25-base Affymetrix probe arrays.

2) Target Concentration. Target concentrations will vary as the gene transcript amounts vary, and the variation extends to the presence of particular exons as well as genes expressed across samples [31-33]. Given the scenario of low transcript (target) concentration, some experimental conditions can be altered to drive the reaction towards duplex formation (increase the rate at which it approaches equilibrium) [14, 15]. The duplex form of the nucleic acids is affected by too high heat, so amplification of the target mass is the preferred method. The challenge is to keep the relative concentrations of individual targets within the mixture identical in the process, which generally utilizes some type of polymerase chain reaction, either as an intermediate or end step [14]. Additionally, 'hybridization accelerators' such as polyethylene glycol (PEG) and dextran sulfate are employed to create a diphasic reagent solution, with the target reagents effectively concentrated in the aqueous phase [34]. This allows the overall volume to be increased sufficiently to cover the relatively wide array surface without diluting the target component. Similarly to the nature of probe sequences, the composition of target sequences may allow the formation of internal secondary structures; in the event that such regions block the probe binding site this will affect the probe's ability to associate with the complementary target region [5, 10, 35]. Target sequences are minimally 1000 nucleotides long (depending on the effectiveness of the molecular biology preparatory processes) and, unless shearing or fragmentation is performed, the

likelihood that no such structural regions form is vanishingly small. Because of the highly folded structures that result there can also be tertiary structures present, compounding the likelihood that the availability of intended sites for probe binding will be sterically hindered, and greatly decreasing the diffusion constant of the molecule in solution (slowing the rate of duplex formation further) [29]. Shearing of targets strands is thus recommended, with the goal of ending with target lengths that are approximately equivalent to probe lengths [5, 10, 35].

3) Temperature. As noted above, the probe length and base composition affect the melting temperature (T_m) of any duplex, where the melting temperature is defined as the temperature at which 50% of the duplex structures are dissociated [22, 25, 26, 36]. Experimental conditions are designed as a best fit for the average of the hundreds of thousands of individual reactions occurring on the array. Solvent components can be altered to change the dielectric constant of the solution, altering the energetics of the product's hydrogen bonds [34, 37].

4) Equilibration Time. As noted above, the competition among highly similar targets for the same complementary probe sequence requires that adequate hybridization time be allocated, in order for the reaction to achieve equilibrium. Short hairpin structures are less energetically favorable than long duplexes, but since unimolecular reactions occur in a much shorter period of time than bimolecular reactions, unimolecular reactions can kinetically trap probe or target in non-optimal forms[5, 10] . The various kinetic properties of the tens of thousands of probes realistically prevent a significant proportion of probe-target interactions from achieving chemical

equilibrium, an underlying assumption for interpreting a Microarray experiment [5, 10]. Achieving equilibrium is a mandate of DNA Microarray experiments in order to ensure a reliable, reproducible, measurement of transcript concentration [5, 10].

Dye-related Factors

After the chip has been incubated with the labeled target/solvent mixture, the unbound reagents are washed away, and the chip is assayed for the remaining presence of the label, which in most cases is a fluorescent dye [13, 14, 16, 38, 39]. The dye absorbs specific wavelength frequencies and emits photons at a second frequency which are captured and amplified by a CCD camera or similar device [40-42]. The readout for the measurement is an image, either from a scanner or actual imager, with varying intensities in 'fluorescent units' at specific locations [14, 15]. Extraction of the intensity at the position known to coincide with a probe is used to infer the type and amount of target present in the duplex. Hence, reliable assessment of a Microarray experiment is dependent on knowing whether the reaction has reached equilibration and how many fluorescent units are emitted per incorporated dye and the number of such dye molecules per target. Dyes may be incorporated at internal positions with modified nucleotides, or may be incorporated onto the 5' end of the target strand through catalytic hydrolysis of the phosphate tail [5]. In either case the enzymes used have preferential incorporation rates with different dyes [40-42]. A common multi-label strategy is co-hybridization, in which a control and experimental sample are labeled with distinct dyes, mixed and hybridized to the same array, with emission filters being used to separate the signals [14, 15]. These two-dye experiments need to include the distinctive properties of the chosen dyes in the experimental design, so as to avoid quenching and dye quantum yield bias in the final comparisons [38, 39]. Affymetrix protocols do not use co-

hybridization and multiple dye strategies, so the procedures developed in this research do not handle this class of problems.

Gene Structure Factors

A probe covers only a small fraction of the possible target sequence, and the investigator should be, but often is not, wary of drawing conclusions about the expression of the 'gene' that, in fact, are true only for a particular isoform of the gene. Current chip designs are attempting to account better for exon coverage and alternative transcript events across sample comparisons by increasing the number of probes and diversifying their placement [9]. Isoform-sensitive analysis does depend on an underlying gene model [43, 44]. While the results of the research presented here are sensitive to probe position and gene models, the resolution of competing models requires wet-lab methods not available to us, so the results shown present likely alternatives rather than absolute outcomes.

Commonly Recognized Confounding Factors

The underlying goal of Microarray technologies is to detect and quantify biological compounds. Any such molecular assay faces the sensitivity challenge of augmenting the signal of very rare events, setting conditions that optimize the specificity of interactions, minimizing background contributions, and respecting device limitations, ideally by including external and internal calibration standards. To date, standard platforms for Microarray analysis have not focused on robust and standardized wholesale technical solutions to these problems, but instead have focused on one particular technical aspect in which to specialize (and segment from the competition) and then relied on statistical methods embodied in algorithmic analysis pipelines for signal

interpretation to cover the remaining aspects [45-49]. Thus, some Microarray platforms implement multiple probes per target assessment in order to improve the specificity of transcript recognition [8, 12]. In this respect, standard ELISAs are more akin to the single-probe-per-gene Microarrays, since ELISAs assay one specific epitope of a protein which is then taken to be representative of the protein in its entirety. However, ELISAs have been modified to compare abnormalities in protein sub-units [50], similar to Microarray exons studies, and the lessons of the suite of controls developed for such assays should be considered by the Microarray standards groups as a conceptual guideline for improved platforms.

Background Contributions

Background contributions to the signal are usually derived at the same time that individual spots are identified and extracted, and algorithms for incorporating the values are often incorporated into the image analysis packages that are optimized for the output of a particular scanner [51]. Common adjustments include subtracting the immediately adjacent fluorescence surrounding individual spots [52] and/or a global adjustment of selected background regions to blocks of spots [52]. Such adjustments assume that any intensity is meaningful: they do not identify the minimal signal that is meaningful as a target response. The best remedy for the uncertainty in the noise level would be the incorporation of a quantifiable nucleic standard(s) on every array, or a calibration standard [7]. Background is contributed from both specific and non-specific sources, and can be truly random (noise, such as from scattered light in the scanner) or systematic (error, such as the effect of SNPs in the sample population) in its contributions to the signal. These effects are dissected in more detail below.

1) Specificity. Probe specificity can be estimated for each set of complementary and near-complementary sequences under the prevailing reaction conditions using

classical reaction equations, modified with nearest neighbor parameters and surface attachment parameters [25]. However, any target genome not exactly similar to the reference genome used in probe design is likely to diverge in unknown ways from the sequence set expected [31]. Thus the precise interpretation of the values is an open ended issue. Common problems that relate to managing specificity include:

a. Cross hybridization. Targets with near perfect sequence similarity to probes have the potential to form sub-optimal but still stable duplexes that compete with perfectly matched targets [29, 53]. The problem is exacerbated with longer oligo probes because the greater stability of longer duplexes can overcome a larger number of mismatches, but it is clear that sequence composition effects can render this a problem even for quite short probes [10]. The heading of cross-hybridization is understood to mean sequences from other sites in the genome than the particular gene that the probe was designed to target, and not variant alleles or alternate transcripts of the intended gene. A complication when the probes are directly attached to the array surface (absence of a spacer) is to determine which of the nucleotides nearest the surface can actually bind the target. Evidence from Affymetrix arrays suggests that the first 6 nucleotides are inaccessible to molecules in the solution phase [10]. This is relevant to the cross-hybridization problem because it results in a shorter string of nucleotides that must be matched, and so a likelier cross-hybridization event.

b. SNPs. Alternate alleles are within-gene sequence variants, and their significance to a DNA Microarray assay is that the intended probe may

demonstrate sub-optimal stability with the desired target [18, 54, 55]. One such class is that of single nucleotide polymorphisms. For short oligonucleotide probes (i.e. 25-mers) a single-base mismatch is supposed to be sufficiently destabilizing that signal decreases markedly [9]. This is by no means consistently observed, since there is a very large context dependence [18, 55, 56]. In the equilibrium equations of duplex formation, the duplex dissociation rate is slower than the single strand association rate, but mismatches dissociate more rapidly than perfect matches; again, the difference in rate constants is dependent on the type and context of the sequence variant [28, 29].

c. Transcript Models. A second class of alternate alleles has to do with transcript splicing. Gene models are routinely redefined as more evidence is published [43], while probe designs are based upon a static definition of the gene model at the manufacturing time. This means that probe annotation must be continually updated by the investigator in order to understand which probes should be deprecated and which should be selectively combined, depending on the desired analysis. Unexpected outcomes are common, including such observations as probes mapping to introns that thereby are expected to have no signal, some of which have high responses, leading to some doubt about the accuracy of the gene model [57].

d. Inaccessible Probes. As discussed above, any single-stranded nucleic acid can form stable self complementary structures, due to intramolecular hydrogen bonding and base stacking [22, 25, 54]. Such structures can

prevent the hybridization of probes to intended (or unintended) target reagents, which will prevent the formation of a measurable duplex product and lead to a false negative result. There may also be increased susceptibility to cross-hybridization if the remaining single stranded target region is long enough to form stable duplexes under the hybridization conditions. Similarly, intramolecular structure in the target may prevent its binding to the intended probe [30].

e. Sample homogeneity. Sample preparations are derived from cell populations in which there is little homogeneity of type and stage [13, 58]. Transcript levels are bound to change according to the cell cycle, environmental conditions, and cellular type [13, 58]. There exist methods of cellular isolation, such as laser dissection and FACS (fluorescently activated cell sorting), or the use of synchronized cultures, that minimize the variation of cell types and localized disease effects, although there will still be variation in the localized environment conditions, cell cycle (unless synchronized), stage of disease progression or response to toxicity effects [58]. Thus, no matter how precisely one is able to target probe specificity to a gene, allele or isoforms, the sensitivity of the experiment relates to the makeup of the sample mixture as a whole.

2) Sensitivity and Scanner Limitations. Scanner limitations were mentioned above with respect to the lower bound of detection, but are important to the upper bound as well. The devices that capture photons saturate well below the limit of response of the Microarray itself [59]. There have been attempts to adjust for this effect by lowering

the energy of the excitation stage, but this raises the lower limit of detection to unacceptable levels [39]. Using two excitation gains successively gives poor results because of photo-bleaching of the dye and the non-linear relation between the gain settings [38]. The effect of saturation is that the scanner makes a stochastic estimation of the relative amount of fluorescence – although a numerical value is supplied the true value should be 'very big' [59]. Scanner manufacturers provide instrument specifications that supply the required sensitivity limits but no pipelines make use of these values. Corrections for individual scanners could be found by the use of either internal or external standards [38], but no 'standard' calibration reagents are available, and indeed almost no one is aware of the problem. Most analysis pipelines do enforce at least lower and sometime upper bounds on acceptable measurement values, but these are typically arise from purely statistical assessments of variance. The better strategy is to remove all data from the stochastic response regions, as in done with ELISAs, and then look at the statistical variation of properly quantified spots.

Microarray Platforms

Published reviews that describe Microarray experiments generally designate two categories: oligonucleotide or cDNA probes [14, 15]. The difference in these array designs is the length of the probe sequence and this difference arises from the production methods. The longer, cDNA, probes are derived from either the excised inserts from clones or (usually) PCR products of those inserts [15]. These probes range from 100 to 1000 nucleotides long and, because of their size (length and tertiary structure), have limited spotting concentration and spot density [5]. Robotic

platforms are used to deposit them at their designated spots. The use of such cloned inserts offers investigators immediate access to reagents and thus ease in modifying chip designs and thereby flexibility in their research. The investment in the robotics is prohibitively expensive to individual laboratories and therefore often research parks and academic centers start core laboratories to provide these services for their community of investigators [60]. The established protocols of these core laboratories present the issue of reproducibility of cDNA Microarray experiments, since quality control and laboratory protocol implementations can vary per core facility [61]. In general our lab does not analyze cDNA array data due to the lack of complete probe characterization and quality control steps in the manufacturing steps.

Parallel to the development of cDNA Microarrays has been the private sector's development of mass-produced, short (25-70) oligonucleotide Microarray platforms [8, 11, 12]. The major companies producing these Microarrays include: Agilent [12], NimbleGen [11], and Affymetrix [8]. The latter two use *in situ*, or on the surface, production of probes by photolithographic methods [5, 10, 62]. The computer-chip derived production methods of these Microarrays allows for extremely high density spotting of probes, and indeed the spotting density has followed Moore's Law [63]. Where Moore's Law has predicted the increase in computational memory over time, the increases in memory are directly proportional to the spotting density of electrical circuits. The differences in manufacturing designs are in the production approaches. NimbleGen and Affymetrix both make use of UV irradiation of reactive subunits at the surface, while Agilent uses piezo-electric delivery of either reactive subunits or pre-manufactured and purified oligonucleotides [8, 11, 12]. There are inherent chemistry limitations to the photolithographic methods, which limits the probe length to a maximal 25 nucleotides, after which the failure sequences begin to accumulate [5]. Affymetrix directly builds the probe on the Microarray chip

surface, while NimbleGen makes use of a polymer tether to build the probes upon [8, 11]. In any design there must be not only resolution between spots but sufficient distance between probes within a spot that they do not interact with one another (essentially a distance defined by the cone that the tethered probe can sweep out [5]. This lesson was learned the hard way by the commercial providers and is often violated by institutional core laboratories, which tend to use very high concentrations of probe solution despite a number of reports indicating more linear responses result from lower concentrations.

Affymetrix U95 Platform

Since the remainder of the research reported here utilized data obtained from experiments that measured transcript levels on the Affymetrix U95 Microarray chip, details of the chip design are included here. The probe design protocol is to perfectly align 25 nucleotide sequences against known transcript regions of the experimental species and identify those that are unique to a single location in the genome [9]. Gene models are matched with 11-16 of such probes, spanning an approximate transcript region of 600 nucleotides at the 3' terminus of the gene [9]. This regional bias reflects the most common method for purifying eukaryotic mRNA, which uses the polyA tail as a capture feature and the basis for the cDNA transformation step. Probes are distributed across this region, but not in a very consistent fashion, since the sequence uniqueness requirement did not allow homogeneous spacing to occur. While such composite or multiple-probe sets may measure the same transcript, it is also possible that different isoforms will be assayed [9].

The Affymetrix company provides a reference Web-available resource (http://www.affymetrix.com/analysis/index.affx) for acquiring information about the probe design

and chip layout [64]. In particular, a master '.CDF' file provides the cell definition, or array design layout, while each sample has a corresponding .CEL file, or intensity output per spot file [9]. The CDF file contains the essential information about the chip's probesets: x and y spotting locations for each probe, the index position of each probe, and each probe's perfect match or mismatch identity. The .CEL files are arranged by the x and y spotting locations and contain the pixel size, signal intensity and standard deviation of the fluorescence across the pixels in the spot, as extracted by Affymetrix's scanner software.

Each probe has a Doppelgänger, a probe incorporating a deliberate mismatch, intended to provide a non-specific hybridization observation [9]. Probes are designed to either have perfectly matched the entire sequence or have substituted a mismatch nucleotide at the 13^{th} position in the probe sequence (homomeric transversions: A ↔ T and G ↔ C). Originally the analysis software from Affymetrix subtracted the MM signal from the corresponding PM signal [9]. The same confounding issues for the perfect match sequences affect the mismatch sequences, although the issue for these sequences is that cross hybridization and SNPs can now attenuate the background correction and thereby potentially nullify the perfect match signal [65]. The confounding factors have made the research community skeptical of the use of mismatch (PM-MM) subtractions and by Affmetrix's own admission, these mismatch sequences are rarely used [66]. There are 409, 600 probes on this U95Av2 array. In addition to the MM probes there exists an additional 3,935 manufacturing quality control probes, all of which lack rudimentary information such as id and/or sequence. These probes apparently are used as controls in the manufacturing process but to the data analyst represent black boxes into the Affymetrix software implementation of spot assessment.

In summary, each probe demonstrates its own binding affinities, and, since on Affymetrix arrays there is no replication of individual probes, variation per hybridization reaction cannot be assessed. Replication is represented solely in the probeset's (transcript level) distribution of signal intensities, and this distribution of the probe intensities then is represented with any of a number of statistical estimators [52]. The aggregation of these probesets into a scalar value representing the transcript concentration can be performed by mean analysis, trimmed mean analysis, median analysis, etc., and there exists considerable debate about the appropriate methodology and whether the mismatch probe information should be used, and how [45, 46, 48]. However, one must acknowledge the dubious nature of a simple interpretation of these 'measurements' for transcript levels when they are only pixel summarizations of fluorescence of sequence subsets of the transcript, where no replication of spots occurs.

Data Analysis and the Impact of the Data Cleansing Methods on Interpretation

The end result of a half-million individual Microarray reactions is the fluorescent intensity of a grid of spots, which the scanner (or imager) records as a Microarray image. Software and algorithms continue to be developed to accommodate adjustments for local and global background spotting intensities and to extract intensities and merge them into scalar gene values [52]. This interpretation of the signal represents a semi-quantitative measurement of the transcript region's concentration [7]. The investigator is then left to make decisions as to how to aggregate individual probes into probesets, including whether to combine probesets that measure the same transcript and/or the same gene [9]. An unexplored aspect of these experiments is

whether analyses based upon individual probes is more or less informative than the aggregated probesets.

Confounding factors have been documented as to their effect upon the aggregated probeset's measurement, and algorithms such as dCHIP, RMA, gcRMA, FARM, etc. have been developed to identify the less variant probes within a given experiment [45, 46, 48]. The assumption underlying most of these algorithms is that cross-sample variance is indicative of biologically confounding factors rather than technical factors. A paradox of this assumption is that these biological factors may be of real biological interest and removing the outliers results in loss of an important experimental result; that is, what if the molecular phenotype of importance is that variation of a set of genes increases as a result of an increase in the presence of a particular gene, rather than that the expression of the affected set all increases (or decreases) homogeneously [67].

The methods reported below were developed to contend with observed short-comings of Microarray results, including the inability to generate cross platform concordance of Microarray experiments, the inability of resultant experimental gene subsets to demonstrate similar performance on independent datasets, and the discordance in resultant gene subsets when different methodologies are applied [35, 44, 57, 68]. In the first part of the research presented here we have been able to show that these experimental discrepancies are readily explained by specific sources of biological and laboratory variation.

The research goal of any Microarray experiment is to identify an informative but still 'manageable' subset of genes, from the thousands of observed genes, whose response correlates significantly with the experimental factor(s) and which can additionally be investigated for

robustness in the predicted expression differences between two given tissue states [69]. These tissue states can be disease conditions, time points in a response curve, drug or toxin treatment groups, etc. Generating such a subset of genes, generally several dozen to several hundred in a list, falls within a type of classical statistical issue now well-understood for Microarray data. That is, any such data set has a large feature (N) to sample (P) bias, where N >> P, in the initial measurement set [70]. In fact, this characteristic violates most statistical analysis method's assumptions about the data [70]. A common test is the identification of unexpressed genes and genes whose expression does not change, such that these genes can be excluded [70, 71]. From here, investigators typically employ additional metrics to facilitate further down selection of the data's features, by identifying those genes with the largest fold-change, noise-to-signal ratio, or some other dimensionality reduction rule in order to identify significant changes [52, 69-73].

A great deal of effort has been extended to refine these identification methods, either in the algorithms themselves or in the statistical power, by exploring false discovering rates, etc [57, 74, 75]. A major presumption about these approaches is that a significant change of any gene's expression levels is found by identifying a large change, whether absolute or as a ratio, compared to a starting level. This does not match what is known about the expression of particular classes of genes, such as those regulating transcription, where very small changes may lead to large effects on other genes [76]. A similar assumption is that the biological tolerance of steady state variance is the same for all genes [67]. The steady state reflects an average of the sample mixtures, but tolerance of expression variation will be different for different genes in different cell types, and in samples having different components [58].

The process of determining a candidate gene list is often multi-step, with a (relatively) simple statistical method being used to obtain an initial down-selected list, in which significant expression change is identified, followed by a more sophisticated technique, in order to suggest a final subset of genes in which correlation to the factor or phenotype of interest is robust [52, 69-71, 77-83]. The assessment of these subsets can be done using supervised or unsupervised methods [69]. Clustering is the most common form of the unsupervised methods, where the goal is to achieve homogeneous clusters [84]. Supervised learning methods develop models from training data and assess the quality of prediction of the test data [77, 85]. Performance metrics are necessary for choosing among the learning algorithms: the most common metric is the area under the receiver operating curve, which incorporates the sensitivity and specificity of the classification results [86]. Other metrics include precision-recall, cost-sensitive analysis, etc. [87, 88].

Specific Aims

A number of confounding factors to Microarray experiments are well described in the scientific literature [17, 18, 22, 26, 59]: to these factors is attributed the relative irreproducibility of Microarray analysis results [35, 44, 57, 68]. While a number of investigators have reported the effect of removing individual classes of contributions on the robustness of the results, to our knowledge no investigation has removed the complete set of factors which we have established in our cleansing pipeline. Of the sophisticated probe cleansing algorithms that have been developed and are commonly used, all proceed by identifying and eliminating probes with large variance, without exploring the underlying cause of that variance [45, 46, 48, 49, 89]. This black box method leads to both inclusion of probes having dubious properties and exclusion of probes that

carry a great deal of biologically important information. The intent of the research presented in this dissertation is to establish a white box probe-based analysis of Microarray experiment results as a rational, scientifically grounded, alternative to current probe cleansing algorithms; to compare the outcome of using that method with accepted blackbox methods, and to analyze the outcomes for biological insights into a particular class of human cancers. Specifically, the goals were to:

1) Design and implement a probe based data cleansing pipeline, in which the contributions of individual factors are interpreted.

   a) Demonstrate that the confounding factors that are removed during the probe cleansing process behave inconsistently across independent data sets, as well as by comparison to their probeset 'mates'. The remaining probesets are demonstrated to have very consistent response patterns across different datasets.

   b) Provide evidence that remaining probeset behavior discrepancies are indicative of currently unrecognized transcript variation events, such as the presence of SNPs or alternate isoforms.

2) The resulting datasets have very consistent expression profiles over many probesets, samples and experiments, and because of this it can be shown that the resulting candidate gene lists are less sensitive to the choice of learning methods. Supervised learning of two independent datasets has been

implemented in order to demonstrate that our method results in data that behaves more consistently upon down selection. Area under the receiver operating curve will be reported for three separate classification algorithms: random forest (RF), k-nearest neighbors (kNN), and linear discriminate analysis (LDA). For comparison, similar datasets have been subjected to the commonly accepted RMA and dCHIP probe cleansing algorithms.

3) The ultimate goal of any Microarray experiment is to generate a subset of genes that either gives insight into a biological mechanism important to the sample state or that gives a high rate of success in predicting the state of a sample. Here we have predicted a set of genes of interest for a two class experiment, normal tissues versus adenocarcinoma lung cancer tissues. This set of genes demonstrates impressive latent structure with and across datasets, as well as supervised classification performance for random forests, kNN and LDA. Comparisons of these genes have been made using the intensity values by the commonly used methods RMA and dCHIP, in place of ours. Finally, the relevance to the biological state of the sample of particular members of the gene list was investigated by literature review

4) Given the world wide incidence rates of non small cell lung cancer and that incidences of NSCLC are increasing in individuals who have never smoked [90], a multiclass NSCLC dataset was constructed from the Bhattacharjee data. This dataset included the adenocarcinoma, squamous cell carcinoma

and normal samples, as cleansed by the methodology presented in Chapter 2. K-means clustering was used to calculate the information gain of the cleansed ProbeSets. The gain criterion demonstrates an efficient methodology to down select to a manageable candidate gene list. This candidate gene lists demonstrates accurate prediction rates for the three disease groups, for kNN and LDA classification algorithms. More importantly what is known about the biology of these genes suggests intriguing differences in cell cycle control mechanisms in the different groups. In particular, the findings suggest the occurrence of aberrant cytokinesis, which may underlie or be resultant in DNA damage elucidating the p53 dependent pathway in the squamous cell carcinoma samples.

Summary

Microarray experiments assess the concentration of a samples' transcript levels at a given point in time. These measurements are the result of a semi-quantitative interpretation of the amount of fluorescence at a given probe spot. The probes at these spots are present in sufficient concentration to drive the expected duplex product formation, but the variation in the response cannot be assessed since most commercial platforms do not provide spot replicates. The measurements of fluorescent intensity are interpreted to represent the concentrations of specific regions of a transcript and for those platforms like Affymetrix that provide multiple measurements over a transcript this redundancy is meant to be equivalent to the spot replicate feature. These platforms have sacrificed measurement replication, ignored scanner limitations, have problematic background correction design and pay inadequate attention to known factors

affecting probe-target duplex formation. Several such factors include the presence of SNPs in the transcript region of interest, cross-hybridization abilities of similar transcript regions having high sequence similarity, and secondary and tertiary structures yielding inaccessible probes and target transcript regions. An additional level of confusion arises from the presence of alternate alleles, whether as uncharacterized SNPs, larger indels and copy number variations, or alternate transcript forms. In order for the results of Microarray experiments to have scientific merit, a significant effort must be made to identify the potential source of deficiencies in probe sensitivities due to any and all of these causes, preferably explicitly, in order to provide more reliable and thereby more reproducible measurements.

Chapter 2: The BaFL Pipeline

Microarray technologies are high through-put platforms that measure some molecular fraction of a sample [1-4]. Gene expression Microarrays assay the concentration of cellular transcripts at the time samples were harvested [1]. Depending on the probe design, the technologies allow one to quantify some fraction of the active genes' transcript levels over the conditions of interest. Accurate assessment of the transcriptional activity depends on how correctly one interprets the source of a signal [5-8].

For example, a number of investigators have pointed out the cross-hybridization problem: many of the probes in any given design do not uniquely bind to a single part of the genome, making interpretation of any measurement arising from such a probe problematic [9, 10]. Elsewhere we have pointed out that probes binding where SNPs are known to occur can result in an altered extent of binding, depending on the alleles present, sometimes with large consequences for the interpretation of the amount of a transcript [11]. We and others have shown that internally stable structures in either the probe or the target that limit the accessibility of each to the other can materially affect the extent of signal [12, 13]. The fluorescent response from the scanner or imager is not consistent over the entire response range of the Microarray itself, so limits must be imposed on the signal range from which the values are analyzed (outside the linear range of the scanner bins must be used instead of fluorescent unit values) [14-16].

It has long been known that the variation due to sample handling may be far greater than the variation due to the primary experimental variable [17], but in the absence of internal controls and general calibration standards we must resort to experiment-specific calibrations [18]. The total fluorescence per array has been previously suggested as one test of batch consistency [19], alternately represented as the average signal per probe or ProbeSet, although those investigators did not incorporate the scanner limitation. This metric reflects the labeling efficiency per molecule, but is not sensitive to sample degradation or large differences in the number of genes expressed, so we extended the metric to include the total number of responsive probes in the linear range [15, 16]. As indicated by the references given for each factor, individual investigators have shown that each of these effects can have a significant impact on the outcome of an analysis, yet, to the best of our knowledge, no one has put all of them together into a simple-to-use pipeline and then tested the final effect on analysis and comparison of experiments.

The significance of the factors varies across datasets by sample characteristics that are independent of the experimental factor (i.e. still biological variation but not correlated to the factor of interest and not subject to controls) and this type of biological variation has created distinct dilemmas for the Microarray field: 1) cross experiment, particularly across platforms, analysis has been deemed impractical and 2) resultant gene lists are not reproducible in classification accuracy across datasets, across classification algorithms, and in their construction [7, 8, 20, 21]. We will demonstrate that the commonly applied statistical algorithms interpret signal intensities differently for each dataset, which changes individual ProbeSet's significance within each dataset. We will also demonstrate that by identifying and removing these types of biological variation the behavior of the ProbeSets becomes more consistent with the experimental factor across datasets, thereby minimizing the identified dilemmas in the Microarray field.

Hereafter the pipeline which we present is referred to as BaFL, or Biologically applied Filter Levels.

Black Box Strategies

A large number of purely statistical approaches have been applied (e.g. dCHIP, RMA, gcRMA) [5, 22-24] to remove variation (sample or technical) unrelated to the factor of interest, but these function as black-box techniques that do not enlighten the investigator about the extent that each factor influences the experimental results. These methods tend to augment the data's sensitivities to classification algorithms; the outcome has been that the processed data performs well within but not between experiments, using the same or different classifications methods [8]. The implication is that these approaches over-train for the factors that apply in one experiment and that those factors are not consistent in the next experiment. This would be expected if some of the result is due to variables with systematic effects on a subset of particular probes, such as the occurrence of different SNP-responsive probes that will give distinct patterns in different study populations [9, 10, 12, 25-30]. In order to demonstrate that the data inconsistencies are sample/population or platform dependent, an investigator needs to be able to delve into the aggregated signal and identify discordant probes and the likely cause of their behavior, and then perform follow-up assays as needed, such as genotyping samples. A black box method does not allow the investigator to understand which particular type of secondary assay must be performed.

Our approach is to identify and remove all problematic probes in a progressive manner, categorizing them as they are removed. Post- BaFL filtering the final set of data from all samples gives a considerably more homogeneous response; in addition the investigator is provided categorizations of the excluded sets that allow examination of each subcategory, and subsets of

probes can thus be reincorporated into the analysis, as the investigator deems appropriate. The impact of each variable is study dependent, but, since our interest is to identify diagnostic signatures, there is a requirement for gene patterns that are robust to individual sampling and technical variation, allowing high accuracy in sample classification, whether binary or multistate [31-35].

An advantage of the multi-probe per transcript platforms is that multiple measurements are available per gene-sample, increasing confidence in the measurement [36]. To retain this feature is important, so, after BaFL filters out probes subject to confounding factors, our analysis pipeline currently requires that a minimum of 4 probes, per ProbeSet, per array, must be present. An optional constraint is that this must be exactly the same 4 probes per array. Variation due to the technical complexity of the assay is completely lab dependent [6, 8, 20] and cannot be quantified in the same way as the factors listed above, so simple statistical tests are used. Two simple tests of overall similarity are: the total amount of response, assessed by measuring the total amount of label present, and the total number of genes contributing to that response, assessed from the total number of probes giving signal. This can be determined at both the probe and ProbeSet level. In these tests, care must be taken only to use signals that can be directly compared, so the manufacturer's specification for the linear range of the scanner is used to set lower and higher bounds for interpretable signal [16, 37]. It is possible that these criteria exclude samples unnecessarily, but in the absence of calibration standards and consistent controls we consider this to be a reasonable, conservative approach [18].

The remaining probes can be collected either as a linked set of values or aggregated into a single transcript- or gene-level value (similar to 'ProbeSet' values computed by other algorithms).

Higher-level analyses are performed on a set characteristic or on the aggregated, transcript-level (ProbeSet) values. The consistency of probe behavior within the remaining set and over the remaining samples can again be assessed. Where differences remain, the investigator has the option of weighting the probes accordingly or classifying the ProbeSet and then handling the separate classes; we have chosen the latter course.

The collection of methods discussed in detail below constitutes a 'white box' approach to data cleansing. They have been instantiated in a software pipeline, with a database backend, that includes the following steps: upper and lower limits on intensities that reflect scanner limitations, elimination of probes with cross-hybridization potential in the target genome along with those whose target sequence no longer appears, elimination of probes sensitive to regions of transcripts with known SNP variations, and elimination of probes with low binding accessibility scores. However, this only removes known sources of variation and therefore there may still be probes affected by uncharacterized transcript phenomena [20]. The BaFL pipeline in conjunction with the ProbeFATE database system allows investigators to identify potential regions of interest, for further analysis.

Materials and Methods

Hardware and Software

A relational database and associated tools system, called ProbeFATE, was used for data storage, organization and simple transformations, which then became the basis for querying for data used in specific analyses. The information system originated as part of the doctoral thesis of Dr. Deshmukh [13], in collaboration with Drs. Carr and Weller and is described in detail elsewhere

(Carr, ms in review). This ProbeFATE system was developed for PostgresSQL 8.0.3 [38] and installed onto an AMD Anthon^{Tm} 64 bit dual core processor running SUSE LINUX ^{Tm} 10.0 as the operating system. Python 2.4.1 [39] scripts were developed with the psycopg2 2.0.2 [40] module to automate the cleansing process and modify the existing system. Through this module data could be extracted and manipulated and analyzed in the R 2.3.1 language environment [41], via the python rpy 1.0 module [42]. Additional software and modules included Oligoarrayaux 2.3 [43] for the calculation of probe thermodynamics and the python MySQLdb 1.2.0 [44] module to enable querying of the public domain Ensembl mysql database [45].

Datasets

Two independent datasets were used in testing the effects of the filtering algorithms. Both were studies of adenocarcinoma patients in which the assays were performed using the Affymetrix AG-U95Av2 GeneChip^{TM}, so consistency of probe placement along the transcripts in the samples is assured. Using this platform, samples are assayed by 409,600 probes across 12,625 defined genes [46]. The largest, or 'Bhattacharjee', dataset (www.genome.wi.mit.edu/MPR/lung) contains measurements taken from 203 snap-frozen lung biopsy tissue samples. The tissues, as described by Bhattacharjee, *et al* [47], consist of 17 normal and 237 diseased samples, including 51 adenocarcinoma replicates, with disease category assigned after histopathological examination. The diseased samples are sub-classified into 5 states: 190 adenocarcinomas, 21 squamous cell lung carcinomas, 20 pulmonary carcinomas, and 6 small-cell lung carcinomas (SCLC) [48]. From this study we used 125 of the 190 adenocarcinoma array results and 13 of the 17 normal results; the selection criteria are described below. The second, 'Stearman', dataset (http:/www.ncbi.nlm.nih.gov/geo/; accession number GSE2514) consists of 39 tissue samples, all replicated, from 5 male and 5 female patients (four samples were taken from each patient: 2

normal looking that are adjacent to the tumor and 2 adenocarcinoma samples); one of the normal samples is missing, presumably for high tumor content. These sample biopsies were harvested using microdissection techniques and then snap-frozen [49].

The most demanding test of a diagnostic assay is whether it is effective in predicting the outcomes of an experiment not included in the development of the diagnostic set. A third, experimentally comparable, dataset was selected from GEO (http://www.ncbi.nlm.nih.gov/geo/): it was published as a meta-analysis of 5 Stage-I non-small cell lung cancers (NSCLC), accession number GSE6253 [33], which will be called the 'Lu' data. Four of these datasets were from published studies of lung cancer, including the original Bhattacharjee dataset. The fifth dataset, also from the Affymetrix^{Tm} HG_U95Av2 platform, consisted of samples from Washington University- St. Louis and included 36 adenocarcinoma and squamous lung cancer patients, all of which were described as being in stage I of cancer progression. This fifth dataset was loaded into the ProbeFATE database system and our probe cleansing and sample cleansing methods were applied as an automated pipeline. The final BaFL cleansed dataset included 10 adenocarcinoma and 15 squamous samples, and 5,311 ProbeSets having at least 4 BaFL-validated probes in common. This dataset will serve for an *a priori* probe selection experiment, based upon the BaFL cleansing of the Bhattacharjee adenocarcinoma and squamous cell carcinoma data.

BaFL Pipeline Components

Probe Filtering

The BaFL pipeline can be divided into two filtering categories, the first, 'probe sequence', category uses only the nucleotide sequence for determining filters, and the second category uses a signal measurement assessment as a filter. The probe sequence filters eliminate probes which

have specific attributes which can be detrimental to the interpretation of the signal intensity including cross-hybridization, loss of target sequence, SNP presence, and structural accessibility. These filters affect all samples similarly. Conversely, the measurement reliability filter affects each sample individually.

I. Unidentifiable Target. The CDF base table for the U95Av2 arrays (Affymetrix NetAffx; http://www.affymetrix.com/products/arrays/specific/hgu95.affx) was queried all 409,600 probes for which the probe sequence annotation was known. There remained 11,432 probes, representing 174 genes, which we eliminated from further consideration. Affymetrix reports the origin/source of these sequences as unknown (personal communication, Affymetrix Technical Support to H. Deshmukh) [13].

II. Cross-Hybridization and Loss of Target Sequence. Probe cross hybridization is the major confounding factor affecting the interpretation of probe responses [9, 10, 20, 21, 25, 50]. We have chosen to follow the Ensembl definitions of cross hybridizations, where 23/25 nucleotides must be in alignment, and we have queried ENSEMBL Biomart (http://www.ensembl.org/biomart/martview/3ee2b94e6eb250f709ffdf9474635fdf) to acquire the list used to perform this filtering step. This process identifies probes that align to a single human genome region, and eliminates those which align to more than one region of the human genome and also those that don't align at all. We note that this comparison is available only for perfect match (PM) probes and therefore if Mismatch (MM) probes are included in the analysis an equivalent list must be acquired and applied, or the level of filtering is not the same in the two categories of probes. Without such a step, incorporation of any mismatch probes (MM) information, as background for PM probes for example, results in a discrepancy in the reliability of the

two measurements being compared. Most investigators no longer use MM values in analysis methods, nor did we do so here.

III. Structural Accessibility. Probe sequences were input to the OligoArrayAux software and the free energies for the most stable intramolecular species were calculated and retrieved [43]. Parameters included: temperature range 41 – 43°C, 1.0 M Na^{+}, and 0.0 M Mg^{2+}. The average free energy across the range was included as probe sequence annotation data in ProbeFATE. This information can be used to remove probes with selected levels of duplex stability. Although there is not a generally accepted cut-off value, we chose a cut-off value of -3.6 kcal/mol as indicative of the presence of an internally stable structure that competes significantly with target binding. In some cases numerical instability (unstable duplex, in effect leading to division by 0 for the free energy calculation) was observed in the output, and such probes were also eliminated.

IV. Presence of SNPs. Probes identified by AffyMAPSDetector as having a corresponding transcript with one or more identified SNPs in the probe-target complementary region (from dbSNP) were excluded [11]. Although the presence of the SNP within a sample may be of particular interest to a researcher, without the individual allele call for each sample these SNPs become a confounding source of variance. For example, the probe may bind strongly to the mismatch instead of, or as well as, the perfect match, and thus the PM value will not reflect the transcript concentration. The current implementation does not extend the SNP filter to the corresponding MM only probes.

V. Measurement Reliability. The individual CEL base tables (i.e. the raw data) may be queried to determine which of the probe signal intensity values fall within a defined range. The defined range represents what is known about the limits of the scanner's ability to provide signal that can be accurately interpreted: above background and

below saturation. The investigator must be aware that signal greater than the higher limit is not the result of an extrapolation in a less responsive but still proportional range (one would see flattening of the curve), but rather a random guess at a 'big' number (one sees greater scatter in the values) [15]. Although the true range is instrument-specific, in the absence of internal calibration controls that let us evaluate this limit we used the range of 200-20,000 fluorescent units suggested by Kachalo, *et al* [15]. An investigator may assign other limits, suggested by experience or available controls, as appropriate. This query can be performed on the reduced probe set, subsequent to the above 4 steps, or it can be performed on the entire dataset and only those probes passing both sets of requirements can be stored for additional analysis.

VI. Statistical Rigor. In these experiments our criterion was that, in a given sample, a particular probeset must have a minimum of four probes remaining, after the steps described above, before a transcript-level value would be calculated (in these experiments the transcript value was the simple mean of the set of remaining probes). Probes in smaller sets were removed. A plethora of procedural choices exists from this point forward. An investigator may choose to simply enforce the minimal acceptable number of probes per ProbeSet and ignore whether the same set is present in each sample, or enforce the complete identity of probes in all samples, depending on the research question. In the results reported here, we enforced commonality of probes. Clearly, the greater the restrictions on number and commonality the smaller the final dataset will be.

In steps I-IV, the probe sequence filters are inherent probe characteristics rather than measurement characteristics and apply equally to all arrays in an experiment done on a particular

platform: only the CDF and probe sequence files are required in order to flag problematic probes. Thus the order of the first four steps is irrelevant and can be set to optimize the computational efficiency. Using our data the cross hybridization filter (II), implemented here only for the PM probes, reduces the dataset most drastically, so if it is applied first the succeeding steps will be accomplished more quickly. Once steps (I)-(IV) have been completed the results are applicable to any future experiments using the same chip design and sequence files. The last two steps described above, (V) and (VI), are experiment/measurement dependent, and it is here that an investigator's choices will affect what appears in the final gene list. Scanner response limits can be re-set in the code, to reflect the behavior of individual instruments

Batch and Sample Filtering

Technical steps will cause the amount of target, the labeling of that target and the effective length of the target to vary independently of the biological factors. Similarly, biological factors, such as secondary infections in cancer patients that lead to dramatic gene expression differences compared to uninfected cancer patients, may obscure the effect of interest. Technical differences tend to be seen in 'batch' effects, i.e. in groups of samples processed in parallel, while biological effects must be screened by comparing an array to the set of all arrays in its class (which may include multiple batches) [19]. The Bhattacharjee data set was explicitly batch annotated [47], while for the Stearman dataset the scan date was used as a proxy for batch annotation: there were 4 dates but in 2-day pairs one month apart, so our assumption is that this reflects only two technical batches. In the following discussion, both individual probe and aggregated ProbeSet values were used to compare individual array to batch and sample class trends, as follows:

I. Probes-per-Sample

a. A filter based on how many probes contribute to the overall intensity, compared to the group mean, using only those probes that survived the above pipeline. Arrays for which this mean exceeded ±2 standard deviations of the group (or class) mean were excluded from further analysis.
b. A filter based on the mean signal per probe, relative to the dataset mean. Arrays for which this value exceeded ±2 standard deviations of the dataset mean were excluded from further analysis.

II. ProbeSets-per-Sample
a. This filter determines how many ProbeSets contribute to the overall sample intensity, compared to the group mean, using only those probes that survived the above pipeline. Arrays where this mean exceeded -1.5 standard deviations of the dataset mean were excluded from further analysis. Samples possessing the lowest surviving ProbeSets were removed more aggressively, since these samples will most limit the population of ProbeSets in the final dataset.
b. The second filter determines the mean signal per ProbeSet, relative to the dataset mean. Arrays in which this value exceeded ±2 standard deviations of the group (or class) mean were excluded from further analysis.

The above two filters were performed in parallel, not sequentially, so there is no order of operations dependence: failing either test was sufficient to eliminate the sample from the pool. Probeset aggregations had the statistical rigor of 4 probes per probeset enforced per individual sample described in the previous section. The filter in IIa is less rigorous than the others, in part because of a desire to retain more samples for the final comparison, accepting that later pruning might be required.

In an independent QC test of the arrays, we performed a parallel analysis of the datasets with the R-Bioconductor `affy` package [19] using mock .CEL files, where probes had been aggregated by batch. The results of this widely accepted algorithm were compared with ours for both batch and sample analysis effects: that is, with and without the 'white box' probe cleansing approach. At each stage of the above-described probe filtering process graphics of the output were generated in order to monitor batch-specific behavior.

Set of probes and ProbeSet Behavior

One goal of this research is to assess the effects that this type of probe and measurement filtering has on the reliability of the set-of-probe patterns across experiments. Data for the first goal was collected using the x and y values to find probes present in all samples, and from there we enforced the requirement that there be at least four probes in a probeset. Set-of-probe profiles (mean and variance per class) are then shown graphically. Next the values in the set are aggregated by taking the average for each sample, which is called a ProbeSet value, despite possible confusion with the Affymetrix ProbeSet [36]. Subsequent down selection used component probe and overall ProbeSets' behaviors, across and between samples, as selection criteria. Within each experiment and for each probe in a set, Welch's t- test analysis ($\alpha = 0.05$) [51] was performed: the variance and sample means in each class were compared and used to define three set-of-probe categories. In the first category, all of the probes in the probe-set demonstrate means which are not significantly different between the sample classes: this class was designated 'Uninformative'. The associated ProbeSets are also 'Uninformative', since the aggregated values must also fail to demonstrate a statistical significance in difference of class

means. The remaining sets possess some number of probes that do show a significant difference in the class means and thus they are considered to be informative. One group contains sets of probes in which every probe demonstrates a significant difference in class means. This maximally and consistently informative sub-group is labeled 'Differentially Expressed' (DE), and here the associated ProbeSet must also show differential expression. The consistency in individual probe and aggregated set behavior is because there is strong concordance in the pattern of differential expression in the probes belonging to the set. The remaining group of informative probes is labeled as the 'Signal' category. In this group, one or more probes have a significant difference in expression from the class mean, that is, does contain a signal, but others in the set do not. Because of this, the associated ProbeSet mean may or may not show a statistical difference from the class mean, depending on the respective values of the component probes.

The underlying goal of the BaFL pipeline is to eliminate probes which may produce variant results because of underlying transcript variation or probe design issues. The resulting measurements should be more reliable and thereby a better interpretation of the mRNA transcript concentration. However, there still exists uncharacterized transcript events which may undermine the signal intensity interpretation [20]. Analysis for differential expression at the probe level allows the investigator the opportunity to identify probes which demonstrate inconsistent measurements with the remaining probe within the ProbeSet. Three hundred and twenty-five ProbeSets were identified to further classify whether uncharacterized transcript events may be occurring. These ProbeSets were considered "Signal' since one or more probes, but not every probe, demonstrated a significant difference in class means. However, after aggregation, the ProbeSet values did demonstrate significant differences in the class means. These ProbeSets appear to represent transcript regions of most importance to the disease state, given the aggregate

is differentially expressed. A python script determined the pattern of expression, $\mu_1 > \mu_2$ or vice versa, and adjusted all probes in the down-regulated class by an increment of 1/50 of each probe mean. This perturbation was decided upon by trial and error. This was sufficient to create pattern inversions in cases of similar expression probes and exaggerate existing pattern inversions, without inverting those probes having differential expression. The ProbeSets which possessed probes demonstrating the pattern inversion after the minor perturbation were reassigned to one of four categories: unique or singular exception, statistical exception, specific transcript region event, and multiple transcript region events.

A Priori Prediction

Candidate ProbeSets were selected from the intersection of the BaFL-validated ProbeSets in the adenocarcinoma stage I and squamous (unknown stage progression) samples in the Bhattacharjee dataset [47]. This set includes 4,257 ProbeSets (from a comparison of 125 adenocarcinoma and 17 squamous samples). Classification results (using kNN, LDA, and randomForest [2, 52-57]), using DE ProbeSets trained on the Bhattacharjee dataset, demonstrated that there is a significant impact of the stage of disease on the expression profiles (data not shown). Therefore, the training set was subdivided, to create a stage I adenocarcinoma group (72), which was intersected with the 17 squamous samples (from multiple stages but not labeled so subdivision was not possible). This yielded 5174 ProbeSets, of which ~4000 were classified as DE, for the (Ad x Sq) comparison. Restriction to the Stage I samples eliminated 16 samples in addition to the batch 3 samples; only 3 of those 16 samples had been originally included in the full adenocarcinoma training set. The intersection of the above BaFL (AdI x SqI) output with the Lu dataset resulted in ~ 400 DE ProbeSets. The constituent probe intensities were recovered, by their x and y location identifiers,

for the 24 samples in the Lu, *et al.* raw data files [33] (filtered as described above), and the probe and ProbeSet presence were then predicted.

Results

The results reported here are divided into three sections: (1) providing evidence for the BaFl cleansing process, including details on the number of probes removed for each probe sequence filter, evidence of sample batch processing variability, and exemplar ProbeSets affected by such confounding factors. (2) the stages and effects of the BaFL probe cleansing pipeline, including graphic representation of each filtering step and the final profile consistency across the two datasets. (3) the research enrichment the BaFL pipeline facilitates, including elucidating potential transcript regions of interest and *a priori* predictions of independent datasets. For the established algorithms we have used the validated samples from the datasets as input, and validated ProbeSets from the author's lists to compare their classification performance to each other and to the results of our method. The classification performance of the author's original gene lists is used to see whether the sample cleansing protocol alone has a significant effect. We show that while previous efforts have identified informative genes, the methods are over-tuned to technical properties (lab specific) or non-primary factor biological properties (such as SNPs), rather than the desired biological response to the principle factor (here, disease state specific).

Probe Filtering Output

The probe sequence-specific filters remove unidentifiable targets, cross-hybridization sources, and probes no longer matching targets, probes having limited duplex accessibility or stability, and those probes for which the presence of SNPs is known. These filters are consistent across all arrays; the independently excluded probe numbers for each filter are presented in Table 2.1. Also presented in Table 2.1 is the total number of excluded probes when the filters are run independently. As discussed above, the ordering of these filters is inconsequential to the final outcome, although some probes may be removed for more than one reason, so the class in which they fall will depend on the order in which the steps are run. The final x and y information for the probe sequences belonging to the intersection of the cleansed ProbeSets is provided in the Supplementary Data, as IntersectionXY.csv. We expect that the linear range limits are the parameters most likely to be changed by other investigators, depending on the type of platform and individual instrument behavior. We have some preliminary evidence that the proper setting of the lower limit can be inferred from the measurement data (shown in the Supplementary Data), but upper limits are more difficult to estimate, and in the absence of calibration standards investigators must continue to rely on the instrument specifications.

Table 2.1: Probe Numbers per filter. The number of probes removed per filter step, when run independently (starting value is 409,600). Values in parenthesis refer to filter effects upon the (201,920) PM only probes. Note that the same probe can be removed for multiple reasons; therefore a simple summarization of probe filter steps does not add up to the number of probes lost over all of the filters in a straight forward manner. For example, if one also considers the probes which were removed for missing sequence information, the Biophysical Filter removes 2.47% of all probes having sequence information, and 2.44% of PM only probes having sequence information.

Filter	**Probes removed**	**% Probes Lost**
Unidentified Target Filter	11,432 (2,836)	2.79% (1.40%)
SNP Filter	7,286 (7,286)	1.78% (3.61%)
Cross-hybridization Filter	246,994 (39,314)	60.30% (19.47%)
Biophysical Filter	21,159 (7,747)	5.17% (3.84%)

Visualization is often helpful in guiding the user to a possible cause for technical problems. In Figure 2.1 we show a virtual array image, generated with the R package `affy` [19], for the Bhattacharjee dataset. This highlights a consistent low-intensity artifact, observed within a small region of the arrays in batch 10 (red circle), affecting ~ 5,600 probes ($2\pi r^2$; radius = 30). To generate the data for this figure, a mock .CEL file for each batch was generated by averaging the intensity of all the probes at a (x, y) location for the samples within an individual batch. Since our methodology constrains the final dataset to consist of cleansed probes common across all samples of interest regardless of the batch, these probes won't be included in the final set of acceptable probes and may lead to the loss of the related ProbeSet as well, if sufficient component probes are removed. Because these probes behave well in all of the other batches, the statistical methods retained the probes but interpreted the related signal for the samples in batch 10 as being significantly lower in expression relative to the other samples, regardless of the disease class.

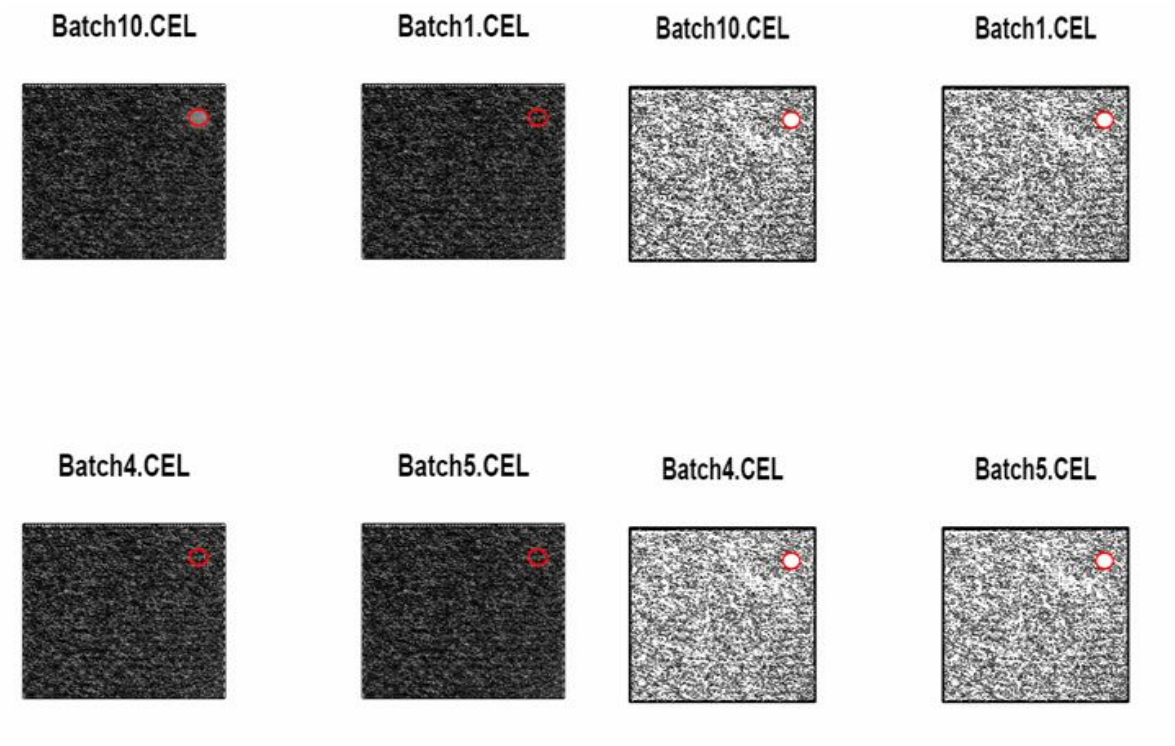


Figure 2.1: Batch images. Image representation of aggregated probe intensities per batch preparation. The red circle depicts a localized feature, most distinct in Batch 10. Left 4 images: raw analysis of samples meeting our filtering criteria. Right 4 images: the same samples with their cleansed probes averaged by batch preparation and only common probes allowed. Probes affected by technical variation are eliminated globally across all batches, including the ~5,600 from Batch 10.

Filter Effects

Since some of the filters lead to a considerable loss of the usable measurement pool, an obvious question is the importance of any or all of the filters. Previously published work has shown the effect that SNPs can have [11]. The largest subset of data is lost from the cross-hybridization filter. It is unclear whether these potential probe-target duplexes form, and how much variation can be expected across a sample population. To investigate this factor, an inversion of the cross-hybridization selection query was performed, which produces a set of probes all of which may cross-hybridize. If these probes are then subjected to all of the other filters we can isolate the impact of this factor. Examples of the results are presented in Figure 2.2, showing how variable

the effect can be. Both the type of pattern and the level of impact differ across individual samples in unpredictable ways: it is not possible to predict particular effects *de novo*, indicating that the filter is important and should remain in the method.

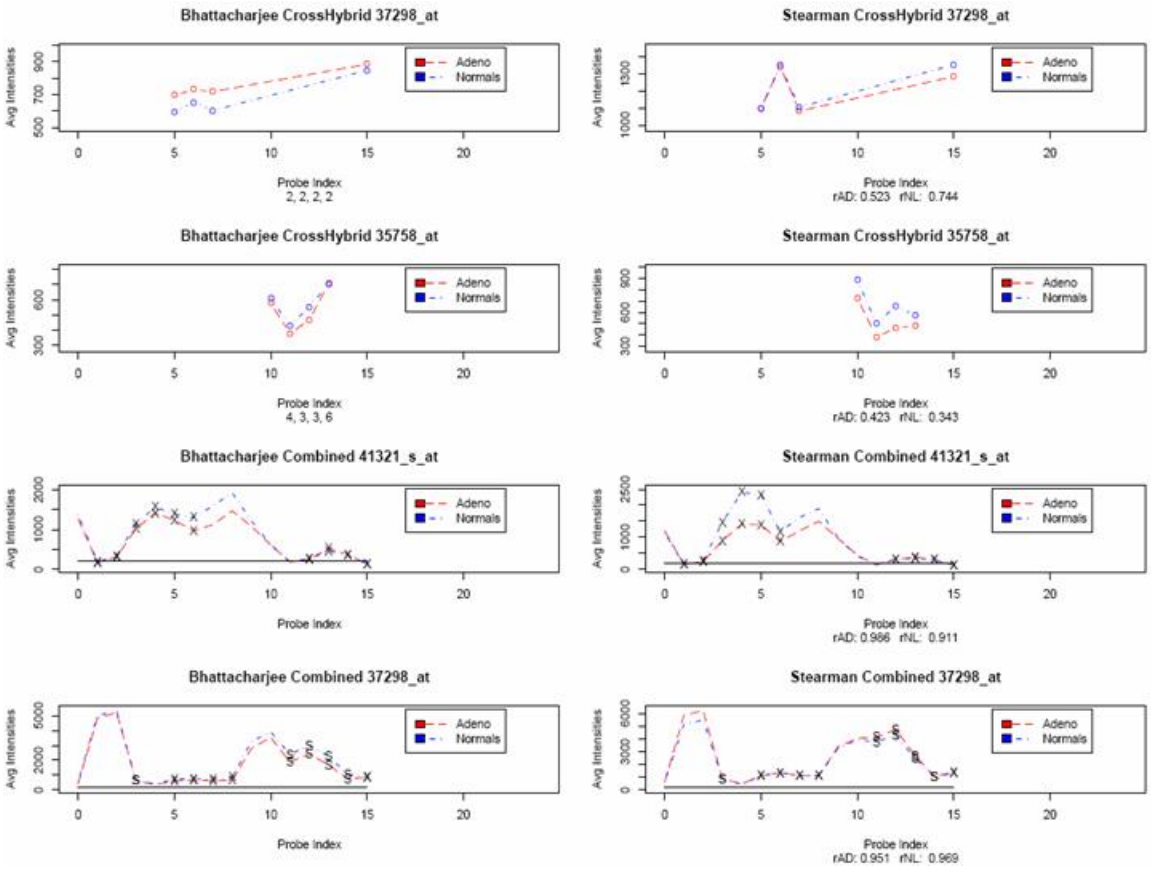


Figure 2.2: Confounding effects. The top 2 rows show the result of selecting cross- hybridizing probes only (other factors still excluded). The bottom 2 rows present ProbeSets in their entirety: at each position the confounding factor is indicated by the symbol (X = cross-hybridization and S = SNP). The binding patterns of these probes (rows 1 and 2) vary considerably between experiments and sample classes. For 37298_at, in the fourth row, we observe that while in this case the cross-hybridization effect is not very strong, especially if averaged into a ProbeSet, 4 of the 5 SNP afflicted probes will produce significant differences in the averages between the experiments regardless of sample class. Probe index 8 for 37298_at (fourth row) failed to meet the linear range criterion in some of the samples (the solid line at 200 f.u.): it also shows cross-hybridization but is omitted from the top row because the other filter also applies. The values for a probe across the sample in the class were averaged to achieve the value shown. Parallel sets of probes, for the two independent experiments, are shown side by side.

Array-Batch Results

Arrays that are outliers due to sample processing problems were identified by comparing individual arrays to the batch-mean values within each experiment. Technical problems are assumed to manifest themselves by increased variance at the measurement level, the tests are

described above. In Figure 2.3, the top row presents the average probe intensities per probe remaining in the cleansed array file (after removing values that fall outside the linear range). In Figure 2.3, the bottom row shows the number of such probes remaining per array, with mean and standard deviation lines provided for comparison. The arrays were grouped in the plots according to their batch membership and are so labeled ('X' denotes batch 10; there is no batch 2 nor 9). Batch 3 as a whole is skewed to the lower end in both tests, so the entire set of arrays was removed from subsequent analyses. Since the probe sequence based filters remove exactly the same probes in all cases, the difference as to which specific probes are removed between the criteria is thus a result of the linear range filter. We believe that Batch 3 suffers from sample degradation since the arrays demonstrate both lower average probe intensity and fewer overall probes. In terms of the sample classes, the Normal samples (in red) were processed across several batches and do not show markedly different overall responses than the disease samples in the same batches.

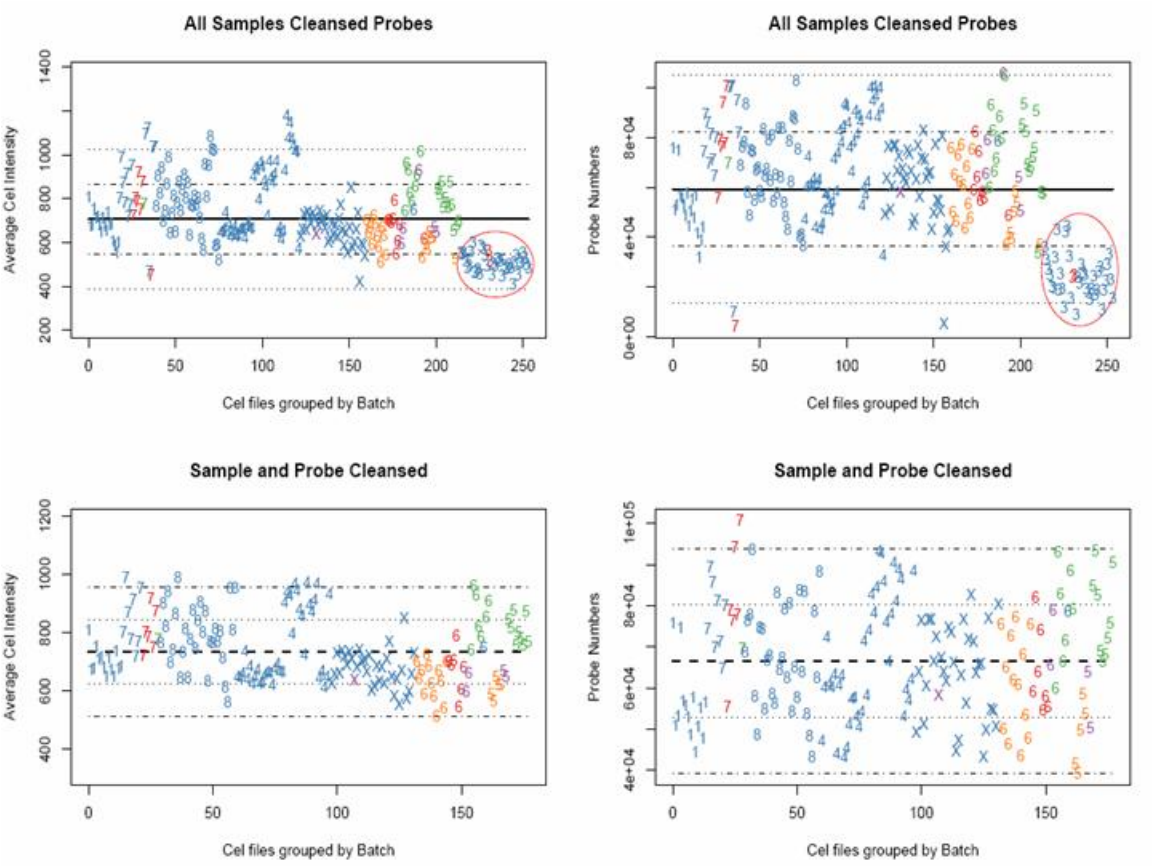


Figure 2.3: Statistical analysis of batches. Graphical depiction of array/batch characteristics in a test of average probe signal (top) and of average number of responding probes (bottom) relative to the mean. The numeral shown indicates to which batch the sample belongs (10 is X), the color indicates disease class (blue=Adenocarcinoma, red=Normal, purple=Small Cell Carcinoma, green=Pulmonary, and orange = Squamous), The heavy line in the middle is the mean intensity; lighter dotted lines are 1 and 2 standard deviation boundaries. The red circle emphasizes the divergent behavior of batch 3 in both tests. There are no batches 2 or 9.

The `affy` package [19] results, when graphed, also indicate that there is a significant difference in Batch 3 properties; therefore this outcome is not an artifact of our probe cleansing methodology. To run this analysis, mock CEL files were created, using the mean probe intensity across the arrays in a batch; this process was followed for the original array files for all arrays, for the original data with outlying samples removed, and for the array files in which deprecated probes *and* sample arrays were removed. Boxplots and histogram densities for the outcomes at

various data processing stages are shown in Figure 2.4. The difference in batch 3 is seen in both the boxplot and histogram densities for the original data, as depicted in the graphs on the left. The middle graphs present the effect of removing all of the arrays in batch 3, but applying no additional sample cleansing steps to the data (used as input into the RMA and dCHIP algorithms [23, 58]). The data distributions still demonstrate a substantial skew. The right-most graphs depict our probe cleansed and sample cleansed data, in log2 space. Note that the linear range cut-off enforces a truncation of the data distributions, which is most visible in the boxplots. The improved normality of the data distribution observed in the last density plot is a result of the combination of probe and sample cleansing; no scaling was applied to the data. Removal of batch 3 accounted for 38 samples.

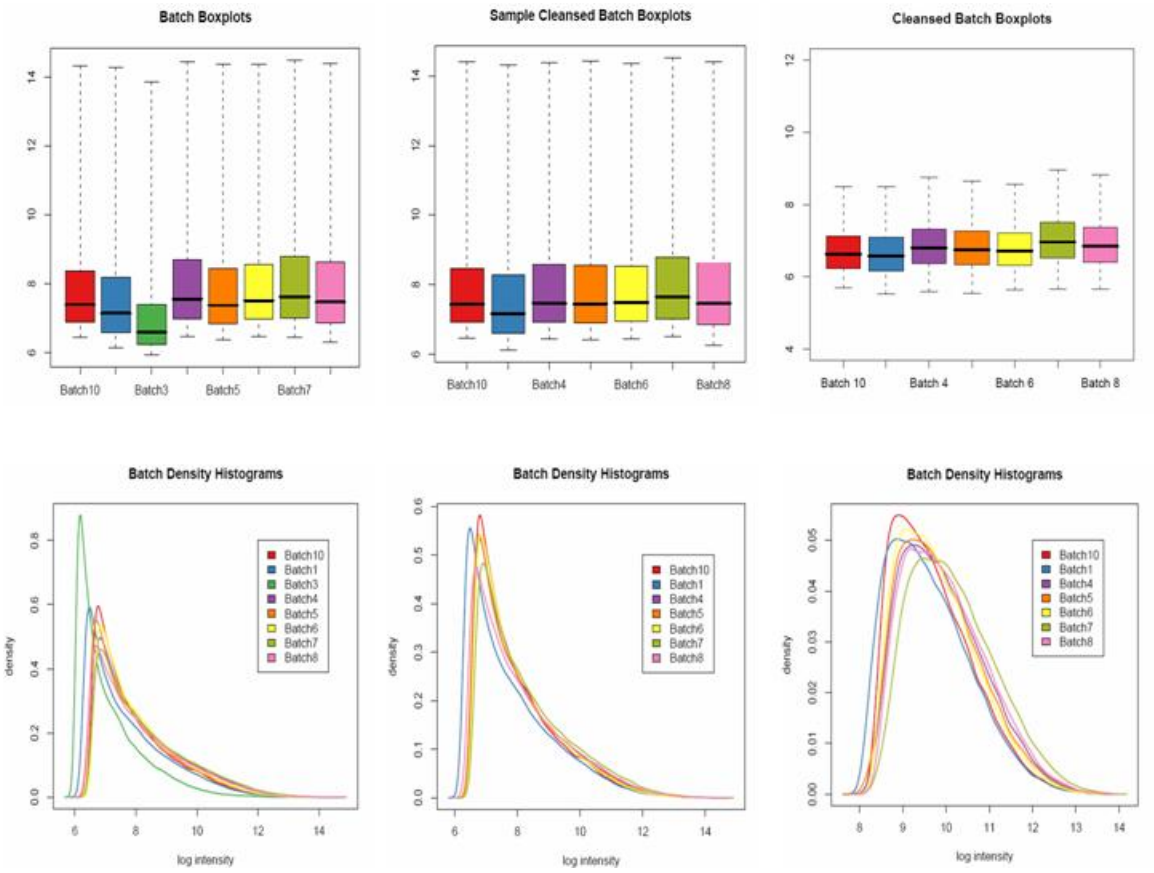


Figure 2.4: Data processing comparisons. Boxplots (top row) and histogram densities (bottom row) of the Bhattacharjee data: summary of batch intensities. The left most pair depicts the completely unfiltered data set, including batch 3: note the obvious offset in batch 3 and the strong skew to the resulting distributions. The middle pair is sample cleansed data, without any of the sequence-based probe filtering, as is used for input to the RMA and dCHIP algorithms: here the skew remains significant but no batches are outliers. The right column shows the output after both probe and sample cleansing methods are applied: note that the distribution is more normal but the tails have been truncated. Also note that the total density scale on the last plot is 10-fold less because so many fewer probes are included (even noise adds up).

In addition to the batch analysis presented in Figure 2.4, the `affy` package provides analyses that are graphed as M versus A and RNA degradation plots. The M versus A plots further support the evidence of a batch 3 effect and are provided in the Supplementary Material, in the Cleansing folder. Conversely, the RNA digestion plot provides no evidence of what seems to be a degradation occurring in batch 3, but given the lack of correspondence between the index and

position of the probe on the transcript, and the many ProbeSets that violate the [0-15] set criterion for assignment of virtual position we do not consider this to be a serious lack of concordance. It is possible that the result is due to a dye incorporation bias [59-61], or a cDNA conversion or PCR amplification problem not assessable by this technique [36].

ProbeSet Assessment Results

Sample filtering excluded those samples which exceeded ±2 standard deviations of either the average probe intensities or the number of contributing probes, at the probe level. The average intensity per ProbeSet and the contributing number of ProbeSets per array were calculated and graphed; the output is similar to what was shown in Figure 2.3 and these plots are provided in the Supplementary Data, in the Filters folder. The ProbeSet number filter was the more stringent of the two filtering steps, even with the exclusion boundary set to -1.5 standard deviations of the mean ProbeSet number instead of 2. This selection was made purposely less stringent in order to enrich the overall ProbeSet intersection prior to down-selection. The final sample numbers remaining for the Bhattacharjee experiment were: 125 Adenocarcinoma, 13 Normal, 17 Squamous, 18 Pulmonary Carcinoma, and 5 Small Cell Carcinoma. The remaining sample numbers in the Stearman experiment were: 17 Adenocarcinoma and 14 Normal.

Consistency of Probe Response

The last refinement in the cleansing process is to identify the intersection of common probes over the samples: the x and y locations were used to identify matched probes across all of the remaining samples. Next, sets containing at least 4 probes were collected, and from these the ProbeSet mean intensities were calculated as the simple mean of the values of the probes remaining in the set, for each sample. Graphical displays of the average probe intensity over the

samples in the class, as well as the average ProbeSet intensity over the samples in the class, show that there is remarkable consistency of the probe response profiles between experiments, some examples of which are shown in Figure 2.5. ProbeSet responses across samples in a class were further categorized based on the outcomes of Welsh's T-tests [51], which were performed, per probe, in $\log_2$ space with an alpha of 0.05. Each set was assigned to one of three categories, described above as U, DE and S. Figure 5 shows examples of some of the results, giving side by side comparisons. As can be seen, in the U class no significant differences in the class means were found, while in the DE category the sets show significant differences (up or down) in the class means. The S category may result in a ProbeSet mean value in either U or DE, depending on the number and contribution to the total intensity of the individual probes in the set. The S category of ProbeSets thus identifies specific transcript regions or probe–target interactions that need additional analysis to be understood. Provided in the Supplementary Materials is the folder ProbeLevel: within the folder are files consisting of 100 randomly selected ProbeSets per classification category along with a random probe QQplot that demonstrates the T-test's validity in discriminating significant differences in this type of data.

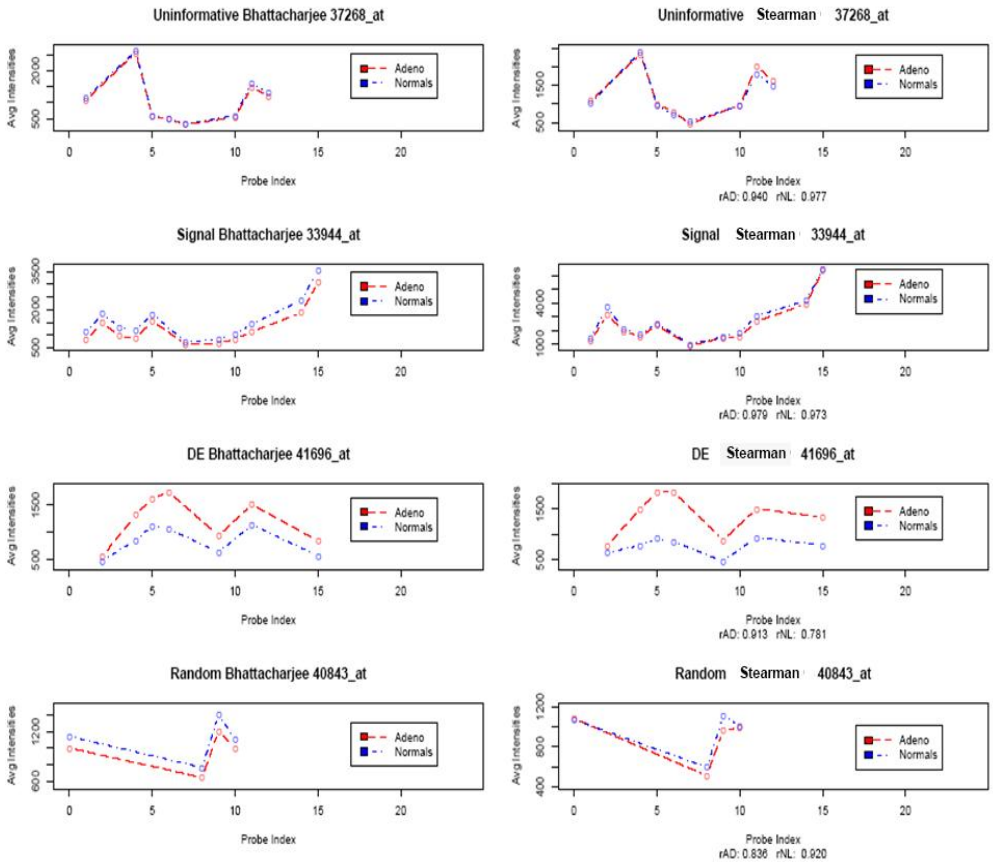


Figure 2.5: BaFL consistency. Demonstration of cross-dataset profile consistency for the three classification categories, as well as a random sampling from the cleansed ProbeSets. Uninformative, Signal and Differentially Expressed (U, S, DE). Left column: Bhattacharjee experimental results. Right column: Stearman experimental results. Intensities are not on the same scale since the labeling was done independently; it is the patterns and relative intensities that are conserved.

Potential Uncharacterized Transcript Events

ProbeSets which were identified as 'Signal', and for which the aggregated intensity demonstrated significant differential expression were identified. There were 325 ProbeSets identified demonstrating the behavior, and any uncharacterized transcript events related to the individual probe inconsistency would be relevant to the disease state. The probes were perturbed by 1/50th of the probe mean to enhance the potential for a pattern inversion and subsequently the ProbeSets were reassigned to one of four categories: unique or singular exception, statistical exception,

specific transcript region event, and multiple transcript region events. ProbeSets containing no pattern inversions were deemed to have a statistical issue, meaning that while the pattern was retained the difference in means failed to be significant and this was assumed more likely to be related to the technical variation across the samples. ProbeSets containing single pattern inversions were separated into a distinct category since no other probe confirmed a transcript isoforms event. These ProbeSets need further individual assessment since neighboring probes may have been removed via the filtering process that would provide more support for such events. ProbeSets with more than one probe demonstrating the pattern inversion were subdivided into a single region event or multiple region events, based upon whether the probes in question possessed overlapping alignments. Examples of each category are presented in Figure 2.6.

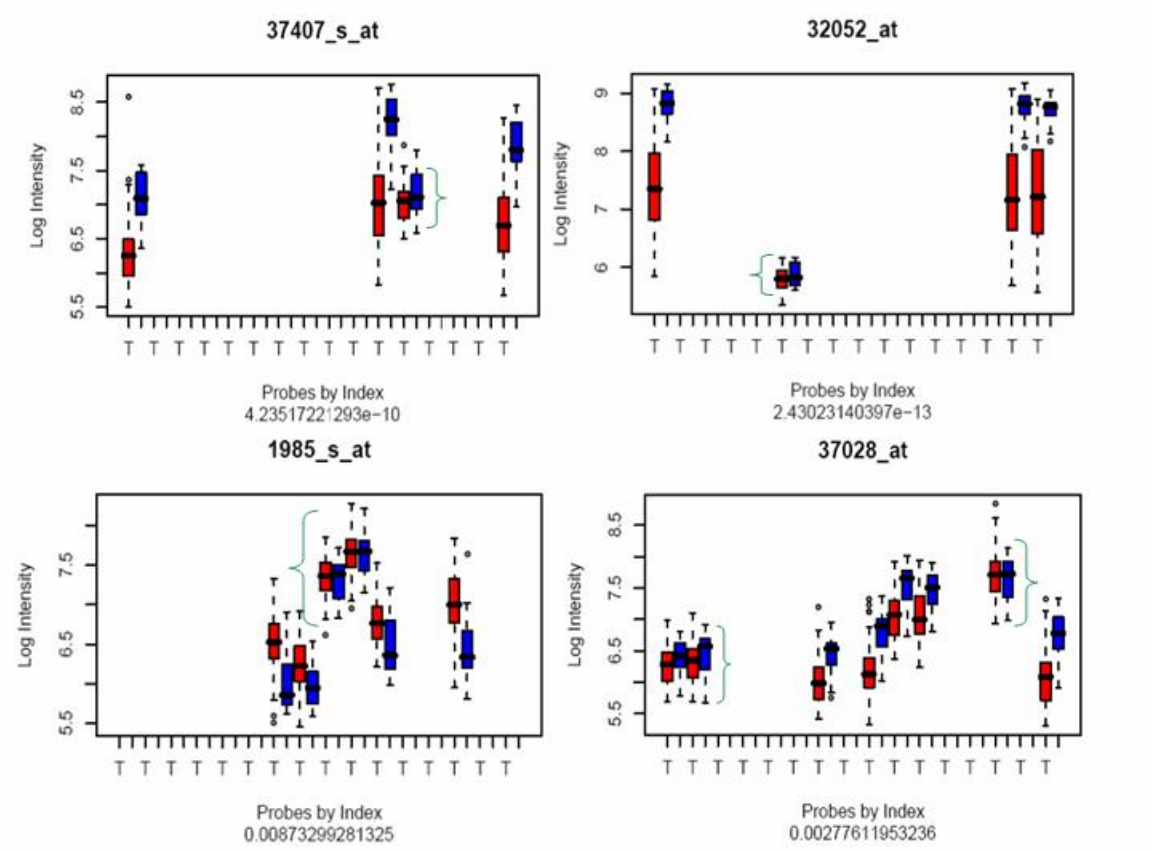


Figure 2.6: Probe-Transcript regions of interest. Examples from the 325 'Signal' ProbeSets, which aggregated as 'DE'. The red boxplots are the adenocarcinoma samples (T) and the normals are in blue (N); average p values are provided below the x-axis label. Probes of interest are indicated with the green brace. Clockwise from upper left: 37047_s_at the 3rd probe in the probe-set failed to demonstrate significant difference in means, nor the pseudo-pattern inversion, suggesting that the difference is a matter of statistical importance. The second probe of 32052_at demonstrates stronger evidence of a transcript event; however an absence of neighboring probes complicates the analysis. The first 2 probes and second to last probe in 37028_at demonstrate 2 distinct regions worthy of additional analysis. The example probe 1985_s_at demonstrates the aforementioned 3 (center) probes as a single transcript region displaying transcript isoform phenomena.

A Priori Prediction

The Bhattacharjee data [47] cleansed via the BaFL pipeline was able to predict the sources of variability in the third Lu dataset [33]; this demonstrates that this is a more robust approach to probe cleansing, rather than either dChip or RMA, which offer *no* such ability. Presented in Table 2.2 are the true positive and false positive prediction rates for the prediction that a probe

remains in the significant pool according to the Bhattacharjee data [47], as observed by the actual cleansing of the Lu dataset [33].

Table 2.2: Apriori predictions. True positive and false negative rates for presence of predicted ProbeSets in the observed BaFL cleansed Lu ProbeSet stage I dataset.

Bhattacharjee BaFL cleansed predictors	**TP rate**	**FP rate**
Adenocarcinomas	99.11%	28.30%
Stage I Adenocarcinomas	98.64%	25.20%
Squamous	74.96%	1.28%
Stage I-Squamous Model	91.64%	13.24%

Table 2.3 presents similar results as Table 2.2, but now examined at the ProbeSet level. This ProbeSet analysis did not include the requirement that identical probes be in the set for every sample, so a follow-on constituent probe analysis was performed. In 25% of the ProbeSets all probes were identical between the Bhattacharjee prediction set and the Lu observation set, while in 20% of the cases the selected Lu ProbeSets contained a complete subset of those in the Bhattacharjee ProbeSets. Conversely, 31% of the probes in the Bhattacharjee predicted ProbeSets were a complete subset of the probes in the observed Lu ProbeSets, and finally, for 22% of the cases, both ProbeSets had at least one unique probe that did not map to the other set.

Table 2.3: ProbeSet behavior predictions. Confusion matrices for classification of the informative nature, comparing the classification of the predicted Lu probes and the observed BaFL cleansed Lu probes. TP is true positive, FP is false positive. The first 3 confusion matrices show the result when t-test classification is done at the probe level, while the bottom matrix is the equivalent comparison performed at the ProbeSet level. ProbeSet level analysis was done for the predictive ProbeSets that are present in the cleansed Lu output (True positive prediction rate for the presence of a ProbeSets was 91.64% and the False positive error rate was 13.24% as shown in Table 2.2).

Level	Confusion Matrix		Class	TP Rate	FP Rate
Probe Level	3904	51	*Noise*	98.71%	12.01%
	77	564	*not Noise*		
Probe Level	540	79	*Signal*	87.24%	1.38%
	55	3922	*not Signal*		
Probe Level	18	4	*DE*	81.82%	0.04%
	2	4572	*not DE*		
ProbeSet Level	4444	13	*Signal* *DE*	99.71%	6.47%

While the characteristic call may have a high level of sensitivity and specificity, as presented above, the influence of a probe within a ProbeSet might either change the overall ProbeSet mean expression and even the direction of relative expression ($\mu 1 > \mu 2 \leftrightarrow \mu 1 < \mu 2$), or make so minor a contribution as to be irrelevant. To test the importance of such contributions, ProbeSet averages were calculated for the probes predicted to be present in the Lu data, as derived from the BaFL cleansed Bhattacharjee Stage I adenocarcinoma-squamous model. Similarly, the BaFL cleansed ProbeSets were aggregated for the BaFL cleansed Lu dataset. Figure 2.7 shows the results of an analysis, for the common ProbeSets, using fold change to summarize disease to normal relationships, for the two sample types in the Lu dataset. The data generated by the BAFL method demonstrates that the aggregation of the predictive individual probes closely mirrors that of the BaFL cleansed probes for the Lu dataset. The consistency lies not only in selecting good ProbeSets, but in handling the constituent probes correctly as well.

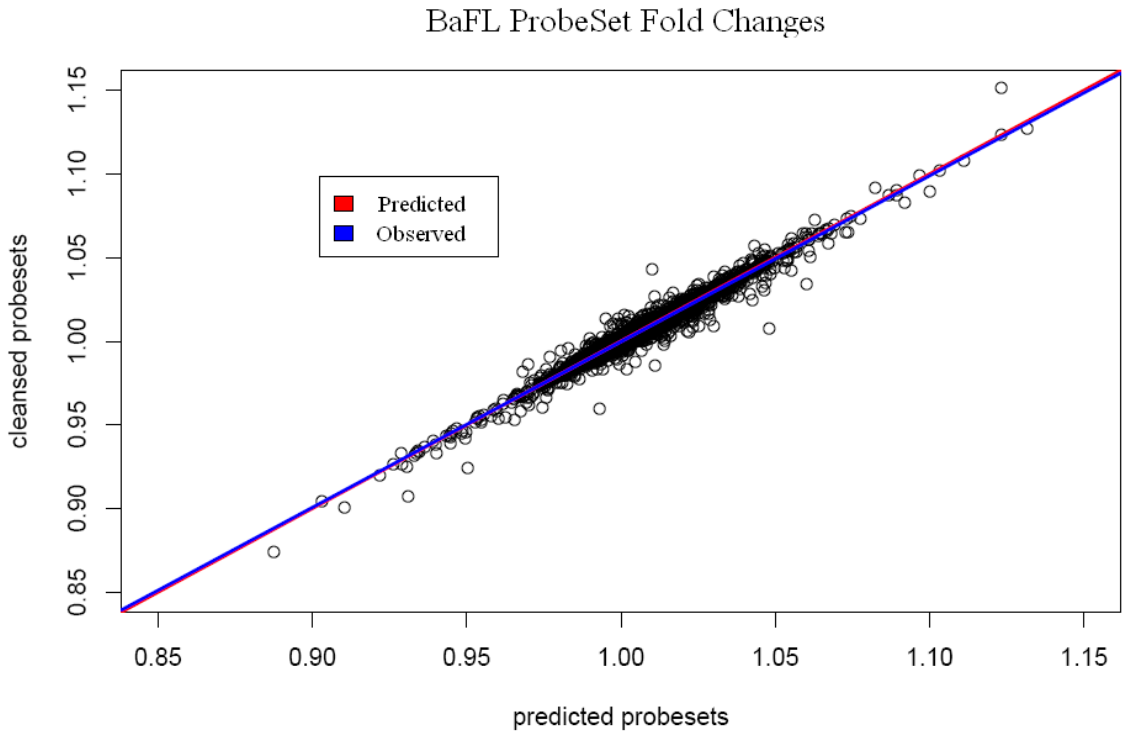


Figure 2.7: Fold change concordance. A comparison of the fold change for our predicted versus cleansed ProbeSets, with a near perfect slope and y intercept.

Discussion

The BaFL filtering process enriches the performance of the standard Affymetrix^{Tm} Microarray experiment beyond that of a single ProbeSet measurement. The BaFL approach allows an investigator to delve into the standard Microarray black box and assess individual probe performance. The ability to evaluate probe performance can facilitate the investigator's identification of transcript regions of interest, which may prove to be correlated to the phenotype of the disease state. Additionally, the BaFL approach allows the investigator to identify entire ProbeSets for which one tissue state demonstrates negligible transcript concentrations in contrast to the second tissue state. Finally, modified CDF files can be constructed to facilitate the

cleansing process. This is demonstrated by an *a priori* extraction of reliable probe and ProbeSets from the third stage 1 dataset, shown below.

Inside the Black Box

The traditional blackbox approach to Microarray data analysis uses a statistical comparison of probes across samples in classes of the experiment at hand, discards (in some cases) or weights component probes according to some 'fitness to a model' scheme, and then aggregates the measurements to give a single ProbeSet value. Thereafter the ProbeSet value is the only factor used as input to machine learning and statistical algorithm development [22-24, 58]. For diagnostic purposes, if the predictive results of these methods are acceptable then the goal has been achieved. However, biological investigators are often motivated by the desire to understand the mechanisms that cause a gene to appear on such a list [7, 8, 20, 21, 50]. Being able to target specific mechanisms may allow an investigator to select a 'discarded' probe for further study: here we are thinking particularly of those probes that are discarded because they respond to SNPs in the coding region, which may in fact be extremely important to the phenotype, if the investigator can apply a follow-up test to qualify the samples. Despite our attempt to identify all such factors, it is clear that we have not done so, since we end up with three response classes and not two in the last analysis stage. We propose that, by doing Welch's T-test at the probe level during the aggregation process, an estimate of the presence of such factors is produced, and the resulting probeset value can be annotated, i.e. affixed with a numerical or categorical denotation (such as our 'U', 'DE' and 'S' labels), based upon the agreement of T-tests results. Uninformative probesets thus are comprised only of probes showing no difference in the means between classes (for a given allowed variation) while the DE probesets are comprised only of probes all of which show a difference in means between classes, as depicted in 2.8. These

probesets can be reliably trusted to demonstrate the same t-test results for the aggregate as the component probes. In the remaining ProbeSets the component probes do not show a consistent response pattern: in the few we have looked at in detail the reason seems to be the probable presence of alternate transcripts, but we do not suggest that this will always be the case.

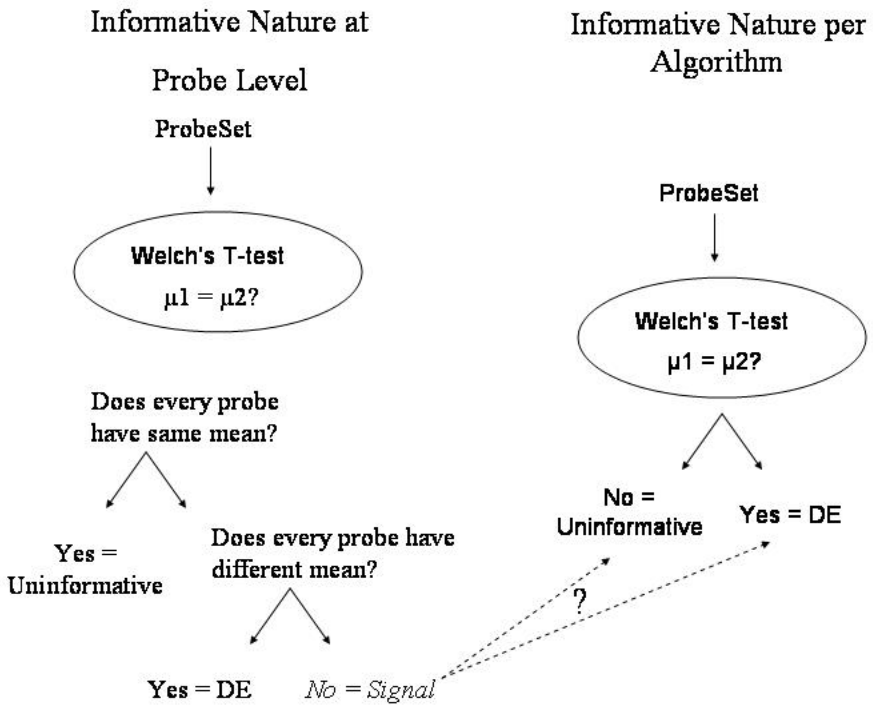


Figure 2.8: Analysis schematic. Schematic depiction of down selection for the white box analysis and the standard black box analysis.

This type of categorical grouping of the ProbeSets facilitates targeted down selection of the dataset, or alternatively a rescue of specific probes if additional assays can be performed. Also, in comparison to the typical blackbox approach, this type of down selection is more stringent, since the criteria for reaching concordance are more exacting. Table 2.4 provides the results of such down selection for the two datasets.

Table 2.4: ProbeSet behavior of probe level analysis. Number of ProbeSets in each response type, for each experiment, and in common, following use of the BaFL protocol.. Note that the 'informative' category is based upon the probe behavior within a ProbeSet, but for the Signal class the two groups may not agree as to which probes in the set differ, a not unexpected outcome if transcript isoforms vary by individual

	Bhattacharjee	**Stearman**	**Intersect**
Cleansed ProbeSets	4,253	6,506	4,200
Uninformative	2,810	1,233	1,225
Informative	1,443	5,273	1,219
Signal	1,288	4,536	937
DE	155	737	79

Transcript Regions Identified by Signal Probesets

Individually, there were 75 ProbeSets failing the statistical power test, 104 ProbeSets having single probes in which a measurement issue, 18 ProbeSets conforming to the single transcript region event criterion, and 128 ProbeSets suggesting distinct transcript region events. More importantly, all 325 probes required additional analysis in order to understand the nature of the information in the signal, and the BaFL approach allows the investigator to assess and prioritize the ProbeSets to evaluate. In addition to identifying the rule under which particular classes of probes were excluded, our method provides a category for sets containing probes with variable behavior (some of which might have been excluded by the two statistical methods, depending on cutoffs of variability chosen). To highlight why this is important we provide two specific examples. For the two ProbeSets (1985_s_at and 39073_at) mapping to the NME1 gene, both were categorized as Signal sets, where not every probe within the accepted probe-set is differentially expressed, while a third ProbeSet that maps to the same gene (1521_at) is classified

as purely DE (differentially expressed). The NME1, non-metastatic protein 1, gene is interesting since it has been associated with metastatic progression in many forms of cancer [62-69]. Closer inspection of the component probes in the two Signal ProbeSets shows that probes specific to a particular region of the transcript are the probes giving discrepant signals. The probe aligning to the 46,593,573rd (start position) nucleotide within the gene gives similar expression between the two tissues states, but as the probes traverse the region, the direction of differential expression (adenocarcinoma > normal) inverts and, finally, the direction of difference is restored with the probe that aligns to the 586th nucleotide. The alignments of these 5 probes are demonstrated in Figure 2.9. These alignments indicate an 18 nucleotide stretch of the transcript that correlates with the discrepancy, which could be evaluated for structural variations in the transcript.

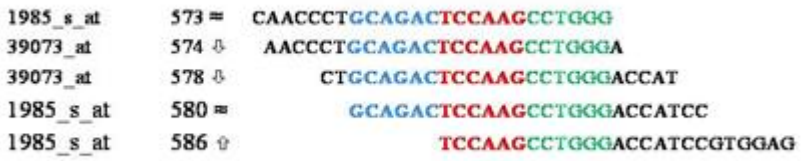


Figure 2.9: Examplar transcript region of interest for NME1. The 5 probe alignments against the transcript, for two NME1 ProbeSets (indices not shown but starting position given). The expression patterns are depicted (disease: normal change) with ≈,⇩,⇧symbols. With the *exception* of the final 1985_s_at probe, those shown have different responses compared to the remaining probes in the three NME1 ProbeSets. The overlapping region (12-18 nucleotides beginning at transcript nucleotide 580) is color coded into 3 sections of 6 nucleotides each. The 6 nucleotides shown in light blue could contain a splice junction or SNP that is not present in the 3rd probe of 1985_s_at which aligns at the 46,593,586th transcript nucleotide. Similarly, the six red nucleotides could represent similar transcript phenomenon, however the presence for which may be sterically masked by the plating properties for the same third 1985_s_at probe, which retains the up-regulated expression levels. The final six nucleotides (light green) have the potential to form a hairpin loop downstream of the red region, thereby disrupting binding of the target in the final probe.

A Priori Prediction

We demonstrated that *a priori* probe prediction is a feasible approach based upon the cleansing results provided by the BaFL pipeline. The probe prediction implemented in this study included the incorporation of the linear range filter. The linear range filter is the only filter which affects samples differently based upon the biological and laboratory variation. Therefore it is paramount that the cross experiment probe extraction is performed upon similar disease studies and similar tissues. This is apparent with the squamous cell cancer data models which generated the lowest probe true positive rates (probes predicted to be thereafter cleansing and actually were). While the false positive rates for ProbeSet prediction for the adenocarcinoma models seem high, we can partially explain some of the false positives by considering that ~5600 probes were removed due to the batch 10 localized bare spot effect, presented in Figure 2.1. These probes would likely not have been removed via the BaFL pipeline for this third dataset. Their presence would therefore be accounted for in the false positive rate (10%, data not shown) observed at the probe level. Removal of these probes may have additionally eliminated what would otherwise have been reliable ProbeSets since the affected ProbeSets may not have met the requirement of 4 constituent probes for the Bhattacharjee dataset. This effect again would be observed via the 25.2% ProbeSet presence false positive rate as indicated in Table 2.7.

Modified CDFs for Computational Efficiency

Since the probe characteristics are universal to an array design, one can easily construct a modified CDF, which decreases the total number of probes that must be considered in an analysis to those that are usable, and thus improves the computational requirements for an analysis. We expect that different investigators will have preferred CDFs: for example, the cross-hybridization

filter acts as a PM only filter and if a mismatch adjustment is wanted then the investigator will perform a preliminary analysis and incorporate this information into a modified CDF.

Sample Cleansing

Sample comparisons are usually performed prior to any data assessment, which can lead to erroneous conclusions about which are the true outliers. We have presented a protocol that proceeds via measurement characteristics to perform batch analyses for technical problems, and follows up with probeset characteristics thereafter to manage biological outliers. The selection of stringency is up to the needs of the investigator: when we relaxed the sample filtering process for the Bhattacharjee adenocarcinoma versus normal samples, an additional 28 samples and approximately 400 probesets were included, but the classification accuracy for the three algorithms suffered (data not shown).

Platform Enabled Analysis Flexibility

Although not the primary focus of this report, the analysis platform we have used provides great flexibility in selecting particular types of probes for detailed analysis. To produce the top half of Figure 2.2 we selected only cross-hybridizing probes, to highlight how different the response pattern they give is to the highly cleansed probes; to perform this analysis required that we alter a single query. Another experiment identified which genes were considered present in the adenocarcinoma state (minimum of 4 cleansed probes) and not present in the normal state (i.e. a plus/minus analysis rather than relative expression analysis). This analysis identified osteopontin: further investigation has shown that it has been implicated in lung cancer development and patient survival [70-75].

Conclusion

We have presented a comprehensive protocol for preparing data for gene expression Microarray analysis, using a suite of probe and measurement based filters, and have shown that by so doing more reliable probe-target measurements result, whose trends are consistent across independent experiments. While individual components of our protocol have been published elsewhere, to our knowledge the methods have not been integrated together and the overall effect assessed. Understanding contributions to a response allows researchers to have more confidence when making cross experiment data comparisons, which will facilitate our understanding of gene behavior within a cell. We do expect that this type of analysis will only be improved with the addition of more sophisticated noise reduction methods applied to data prepared in this manner. Finally, probe based analysis is greatly simplified if carried out with a database system such as ProbeFATE, which uses the probe as the atomic unit and has been optimized for manipulations and aggregations that build specific subsets based on user-coupled criteria.

Chapter 3: Down Selection

The focus of this chapter is to compare the performance of the BaFL method's interpretation of signal intensity measurements, in the form of aggregated values called ProbeSets, against the analogous ProbeSet interpretations of two statistically based algorithms, RMA and dCHIP [1, 2]. The test used in the comparison is the performance of each method in cross-dataset classification experiments. RMA and dCHIP (here, the R implementations thereof) are two commonly used methods that, like BaFL, reach down to the probe level values in order to determine which members of a set of probes are to be included in an aggregated ProbeSet, although they differ from each other as to how the remaining members of the set should be weighted in that aggregation [1, 2]. One challenge in doing this comparison is that RMA and dCHIP are, at the measurement level, black boxes, in that the user does not know, and cannot retrieve, those probes that are left out (or included), nor is the reason for elimination explained That is, although the algorithms are fully accessible, they were not designed to assess or report on probes at the individual level but only as aggregates. On the other hand, for the BaFL method we can precisely assign the reason for eliminating a probe, extract information as to whether additional reasons for eliminating a probe exist, and show what would happen were it to be included; for the other methods we can only compare the ProbeSet differences and effects on subsequent analysis outcomes. In this chapter we perform parallel analyses and compare the outcomes, and trace back the root of the discovered differences to the extent that the algorithms permit. In addition, a comparison is made to the 'significant genes' lists of the original authors [3, 4], bearing in

mind that they are most aware of the experiment and the conditions prevailing when the measurements were collected.

Materials and Methods

Data

In the following comparisons, only those *samples* that were judged acceptable, using the sample cleansing and batch comparison methods described previously (see Chapter 2) were used in any analysis. This means that the base dataset used in the comparisons is consistent across experiments (138 Bhattacharjee samples (125 disease, 13 normal) and 31 Stearman (17 disease and 14 normal) samples, see Supplementary Materials for the file names and files. Each of the three algorithms was used to generate ProbeSet values in these samples (signal intensities were transformed to $\log_2$ space). While the RMA and dCHIP algorithms yield the full set of 12,625 ProbeSet values, the BaFL protocol retained 4,253 Bhattacharjee ProbeSets and 6,506 Stearman ProbeSets, with 4,200 of these being concordant ProbeSets. That is, in addition to the ProbeSet values themselves, the primary difference in the input matrices tested for differential expression is the number of ProbeSets considered. This affects the number of candidate genes in subsequent lists that are available for model construction in the final classification analysis. Two types of comparisons of the three probe cleansing methodologies are made, following a typical analysis framework: down selection and validation of a candidate list. The first analysis is based on the down selection to significant features as provided by Welch's t-test [5], and the validation of candidate gene lists is considered with respect to those of the original investigators proposals [3, 4].

Data Analysis Overview

1. Each of three probe-cleansing methods (RMA, dCHIP, BaFL) is used to generate ProbeSet values, on the same sample sets from each experiment (2-state case).
2. Down selection was performed for each cleansing methods' interpretation of the data. Down selection yields the identification of differentially expressed genes, starting with the ProbeSet values produced by each method, based on the outcome of a Welch's t-test of those values across the sample sets [5]. Three such down selection lists are generated: one list of DE genes from each experiment and a third that is the intersection of those two lists. The values of the genes in the lists (Stearman DE, Bhattacharjee DE, and Intersection of DE) then are used as input to three types of classifiers; kNN [6, 7], LDA [8, 9] and RF [10-12], and the resulting models are assessed, based on the AUC curves [13, 14], for their cross-experiment sample class prediction ability relative to the base model (ALL), the complete set of genes' values.
3. A second type of comparison uses the two candidate gene lists proposed by the Bhattacharjee, *et al*. authors [3] and the candidate gene list proposed by the Stearman, *et al*. authors [4], sub-selected in each case for those genes passed by the BaFL pipeline (but not necessarily identified as DE). A fourth candidate gene list comprised of the BaFL-passed and intersecting t-test identified DE genes for both BaFL datasets, and is the same final list which was used in step 2. The four lists used the ProbeSet values originally suggested by each cleansing method, (not the values that resulted from the methods of the original papers, since the underlying sample sets have been modified) and then proceeds as in step 2 for a comparison of classification strengths based on the three types of models.

These steps are discussed in more detail in the following sections.

Probe Cleansing Methods

The BaFL protocol (Chapter 2) was applied to the datasets and the mean of the cleansed probe values was computed to obtain a scalar for the remaining ProbeSets in each sample; as before we required that there be 4 or more probes in a ProbeSet that were common to all samples in the class; here we look at a two-class problem. RMA and dCHIP implementations in `R-affy [15, 16]` were used to generate ProbeSet values for the same set of samples. Li and Wong proposed dCHIP as the implementation of a model based expression index (MBEI) [2, 17]. The heart of their algorithm utilizes a weighted average of mismatch differences:

$$\widetilde{\theta}_i = (\sum y_{ij}\phi_j) / J, \quad (1)$$

with i representing the samples, J representing the number of probes for a probeset y_{ij} and ϕ_j the probe level mismatch difference. The weighting scheme favors probes with the largest PM/MM difference. The workhorse for this algorithm is the probe sensitivity index (ϕ_j), which identifies probes with large standard error and negativity (MM > PM) [2, 17]; the likely sources of this effect are cross hybridization and laboratory handling. However, these sources are going to be inconsistent across experiments. Their publication makes the claim that the probe sensitivity index should be independent of the tissue type [17]. One can readily give a counter-example to the assumption of this statement: given a probe that has two potential transcript isoforms to which it can hybridize and 3 tissues, if tissue A has little expression for both transcripts, tissue B has expression of only one transcript and tissue C expresses both transcripts well, then the probe sensitivity index will reduce the contribution of the probe consistently only for tissue C. RMA implements an additive model and considers only the PM data, after performing a background

transformation and quantile normalization of the data by array [18]. The additive model is as follows:

$$Y_{ij} = \mu_i + \alpha_j + \varepsilon_{ij}, \quad (2)$$

where i is the samples and j is the ProbeSets. This model averages the (PM only) probes per sample, accepting some random error in the model (ε_{ij}), and assumes that the probes have been designed such that the accumulated probe affinities, $\alpha_j = 0$ [18]. The algorithm implements the median polish [16] to detect outlier probes, which violate the probe affinity assumption of the model. In both of these algorithms, some ProbeSet value is determined for every set on the array. In contrast, the BaFL cleansing protocol eliminates probes and through enforcing a minimum set size (for statistical rigor) and consistency of set members across samples, may result in the removal of entire ProbeSets. This often results in the absence of a large fraction of the original data set: in the case of the Bhattacharjee dataset 66% of the original ProbeSets are removed. It must be acknowledged that the disparity in the number of genes in the input set makes a straightforward comparison of the output lists of the three methods problematic, but it is possible to highlight some sources of error that the statistical methods do not identify and exclude. The output data files are included in the Supplementary Materials, in the Data folder

Down-Selection

A Microarray experiment typically has 10,000 features to explore, at the gene level, of which around half are expected to be expressed; generally only a small proportion of the genes will be differentially expressed. Many of the statistical methods used to determine whether differences

are significant assume such a data structure [7]. This still leaves several hundred values to compare per sample. For classification purposes a much smaller diagnostic set is desired, so analysts typically employ a down-selection method to their dataset [12, 19-29]. The rationale for such an approach is two fold: it should improve the statistical power of subsequent analysis, by minimizing the N>>P issue and it will enrich the impact of the final candidates, by eliminating weak classifiers within the candidate list [11, 12, 20, 21, 27]. Weak classifiers are features within a solution subset which are not critical to the subset's classification performance [12, 30, 31]. Common approaches to down selection in Microarray analysis pipelines have included: significance analysis of Microarray (SAM), t-tests, fold changes, expression differences, signal to noise ratios, etc. [12, 21, 23, 26, 27, 32]. For the following analyses we employed Welch's t-test on the ProbeSet values, rather than the individual probes across the ProbeSet since this is the only comparison possible with the RMA and dCHIP methods (see Chapter 2 for details). Welch's t-test first requires performing the F test for equality of variance, as given by:

$$F = \frac{s_1^2}{s_2^2}, \tag{3}$$

Welch's t-test is formulated as:

$$t = \overline{X}_1 - \overline{X}_2 / \sqrt{s_1^2 / n_1 + s_2^2 / n_2}, \tag{4}$$

where $s_1 = s_2$ given equal variance, which must be tested for. The variable $\overline{X}$ represents the sample population mean, and n is the sample size [5]. The t-test is usually considered to be a weak diagnostic and efforts to control the family wise error rate through Bonferroni corrections or

to control the false discovery rate are employed to improve the down selection process [33]. However, in the following experiments we did not implement such correction efforts, or other algorithmic optimizations, in order to see the full range of possible solutions and to see where overlap of genes occurs even if the significance assigned by the different methods is quite different. The Bonferroni correction has been applied to the datasets in Chapter 5. The results of the t-test categorized the ProbeSets into two classes: uninformative and differentially expressed. The datasets are available in the Supplementary Materials, in the Data folder. Additionally, within the Data folder is a README file, which details the naming convention and the flags which identify ProbeSets and their classifications, including the Bonferroni correction classifications.

Published Candidate Gene Lists

In the original report of the Bhattacharjee experiment, two ProbeSet lists were developed, based upon the signal intensity reproducibility across 45 adenocarcinoma replicated samples, where reproducibility was assessed based on whether these ProbeSets had Pearson correlation scores above the 0.8 and 0.85 thresholds [3]. These lists consisted of 675 and 363 ProbeSets, respectively, for the correlations 0.8 and 0.85. The cleansing methods used by the original authors were quite complex, and neither the ProbeSet values nor the exact gene lists could, in fact, be perfectly replicated using the published descriptions [3, 34]. We decided to take those genes from their lists that survived the BaFL pipeline, that is, that had been cleansed of the known extraneous factors leading to variability, and see if the genes remaining had strengths for classification that our own down-selection method missed. That is, the point was to determine whether the original selection methods had merits that our own procedures lack. Of the original ProbeSets lists, 267 and 136 (for 0.8 and 0.85 thresholds respectively) survived the BaFL

cleansing process and were assessed for their classification ability. The original report of the Stearman experiment reported a list of 409 genes, which demonstrated concordance in $\log_2$ difference (tumor minus normal) between the murine and human model [4]. Surviving the BaFL cleansing process from this list were 178 ProbeSets. The concordances between the original three lists are relatively small, containing 58 and 34 ProbeSets, respectively, for the 0.8 (Bhattacharjee x Stearman) and 0.85 (Bhatacharjee x Stearman) lists, so the size of the concordant lists against BaFL is not unreasonable. We compared the classification abilities of these three lists to that of the 325 ProbeSets whose aggregate was classified as differentially expressed for both the BaFL cleansed Bhattacharjee and Stearman data. ProbeSet values were again generated by each of the three data cleansing pipelines as part of the comparison. Complete ProbeSet lists are given in the Supplementary Materials, in the Data folder.

Classification

The classification performance of supervised learning methods was assessed, using the various candidate gene lists as training sets. The area under the receiver operating curve (AUC) was the performance metric for all the classification experiments [13, 14]. For each algorithm, the base model included all of the original surviving ProbeSets, either the 12,625 (for RMA and dCHIP) or the intersecting 4,200 (from BaFL). A requirement of the classification experiments is that the same ProbeSets need to be present in both datasets. This only affects the BaFL data because RMA and dCHIP give complete gene value matrices [2, 18]. Therefore, for the BaFL ProbeSet lists, subsets that consisted of candidate gene list intersections were used. The values in the resulting ProbeSet lists were then used, in turn, to train three different classification algorithms: k nearest neighbors (kNN) [7, 23, 35], linear discriminant analysis (LDA) [8, 9, 20] and random forest (RF) [10-12]. The R implementations of these algorithms were used [16]. The parameters

for kNN were (k=3, l=2, with the Euclidean distance), and default settings in R were used for RF and LDA [16]. We chose to use three different methods in order to explore whether the classification performance was specific to the classification algorithm, and these three were selected specifically because they are the most commonly cited in Microarray analysis papers and because their performance requires minimal parameter tuning [6, 11, 12, 20, 21, 23, 27, 28, 34, 35]. Linear discriminant analysis attempts to find the linear combination of features which best separates the data into their distinct classes, by weighting the features based upon their ability to separate the classes [8, 9, 20]. Conversely, kNN and RF classify samples based upon the characteristics of closely neighboring samples [6, 36]. The entire ensemble of features is utilized for the kNN algorithm while RF stochastically builds forests of classification tress based upon the strongest classifying features [10, 11]. After training with values from one experiment, the models were used in tests against the other experiment and the performance was assessed: that is, the Bhattacharjee gene lists were used for training and then the models were used to predict the Stearman sample classes, and vice versa, for each of the types of gene lists described above [6, 36, 37]. This led to 9 comparisons in which the Bhattacharjee data were used as the training set (RMA, dCHIP and BaFL cleansing post t-test, against 3 types of models) and 9 comparisons in which the Stearman data were used as the training set.

The same classification algorithms were invoked for the comparison of the author's lists to the purely BaFL-derived list of 325 DE ProbeSets. This set of experiments is designed to be similar to that of the validation of a final candidate list. Here, we compared the 325 BaFL intersecting DE ProbeSets, to the BaFL-allowed ProbeSets in the author's published lists. Validation of a candidate list necessitates perturbing the designed models over iterative analysis to approach a reliable performance metric [6, 36, 37]. Perturbation of our models was done through random

sample selection, with replacement, for 100 iterations to approach a reliable AUC performance metric [6, 13, 14, 38]. Random sampling in such a manner permutes the data in a fashion that leads to some artificial replication and may omit some samples [12, 30, 31].

Results

The results are given for each of the probe cleansing methodologies as independent analyses of Microarray experiments. There are two major sections: the model's performance before and after down selection, and the validation of a candidate list. Here the candidate list is the 325 intersecting differentially expressed ProbeSets as interpreted by the BaFL process and this list will be compared to the author's pre-existing published candidate lists. Independent validation is used throughout to demonstrate that the BaFL probe cleansing algorithm facilitates cross experiment analysis.

Down Selection

Table 3.1 summarizes the number of DE ProbeSets per probe cleansing algorithm that was the result of the t-test for significant differential expression. The number of DE genes predicted for the Bhattacharjee dataset by the BaFL pipeline approaches the expected number (4,200 * .05 = 210) [5, 33]. This is not true for the Stearman data, likely due to the greatly diminished size of the sample.

Table 3.1: Down selection numbers. Number of ProbeSets giving values in the down selected gene lists that result from applying Welch's T-test to the output of each probe cleansing method, per dataset, with the Base Model giving the original size of each dataset. Each of the down-selected lists is then used as input into the 3 types of models. All of these data are provided in the Supplementary Material Data Folder.

	RMA	**dCHIP**	**BaFL**
Base Model	12,625	12,625	4,200
Stearman DE	5,291	5,208	3,344
Bhattacharjee DE	6,595	6,429	*480*
Cross-Experiment Intersection ∩DE	3,761	3,407	325

Figure 3.1 presents the *p*-value kernel densities resulting from each of the probe cleansing methods, for the Bhattacharjee data and normally distributed random sampling. The top left graph estimates the probability distribution (default Gaussian smoothing) for all the *p*-values for the 3 methods, with respect to the random population. The top right graph presents the quantiles for the three methods, with respect to the random population. While we observe that the RMA and dCHIP kernels appear to be more normal, they also demonstrate a large, exaggerated skew and the quantiles deviate from the expected. The skewed tail represents the population of null *p*-values, as shown in the lower left graph, with the accompanying quantiles presented in the lower right graph. This disproportion of the RMA and dCHIP t-test hypothesis testing results is associated with the skewed batch intensity distributions we presented in Chapter 2. In stark contrast, the BaFL *p*-value density demonstrates a skew for the upper quantiles and becomes more pronounced for the null *p*-values [39]. These graphs explain the observed weakness of the t-test for RMA and dCHIP and are intriguing for the BaFL hypothesis tests, since there seems to have been an increase in the power of the t-test. Is this increase due to the BaFL cleansing process or as a result of the bias in the dataset?

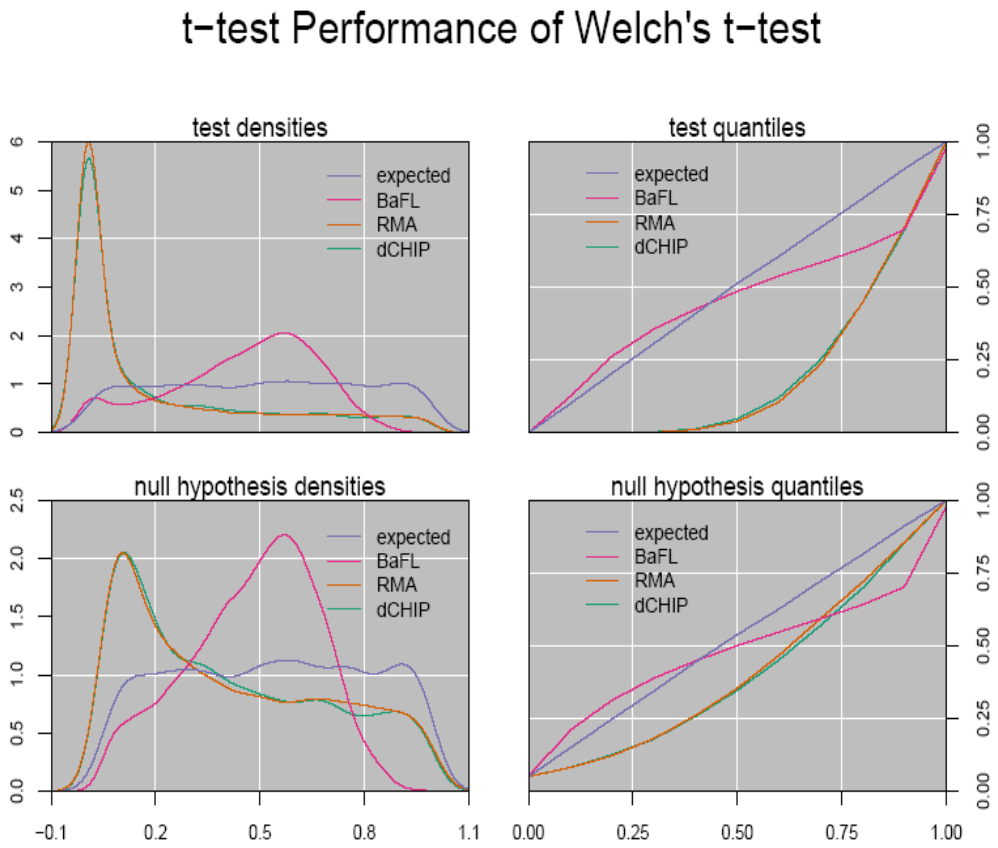


Figure 3.1: P value distributions. *P* value density distributions and quantiles for the Bhattacharjee data as calculated by the Welch's t–test for each of the probe cleansing methodologies, on comparison to normally distributed random sampling. The top left graph presents the Gaussian kernel estimation for the entire *p* value distribution, with the corresponding quantiles presented on the top right. The bottom demonstrates kernel density distributions and quantiles for the null *p* values (greater than 0.05).

Models and Class Predictions

Figure 3.2 shows the sample class predictions when the training set values came from the Stearman experiment. The input ProbeSet lists are labeled according to the description of their generation, and include four categories. The test set was the Bhattacharjee experiment (just doing two sample classes), for each of the three classification algorithms: kNN, Random Forest, and LDA. The last column summarizes each selection level of the data, across the three classification algorithms. For the top row of graphs in Figure 3.2, the number of input ProbeSets in each of the

four assays is given in the second column of Table 3.1. That is, the first input gene list, ALL, is the complete set of 12,625 ProbeSet values generated by RMA for the Stearman samples. The second input gene list includes those 5,291 ProbeSets identified as being differentially expressed in the Stearman sample, the third input gene list includes those 6,595 ProbeSets identified as differentially expressed in the Bhattacharjee dataset and the fourth candidate gene list has the 3,761 genes identified as being in the intersection of the second two lists. The second row of graphs uses the ProbeSet values generated by dCHIP, which are different from those given by RMA, and this leads to somewhat different candidate gene lists (see Figure 3.7 in the Discussion for more details) but follows the same pattern of assays, with the numbers given in Column 3 of Table 3.1. The third row of graphs requires more description because the 'complete' matrix of ProbeSet values generated by the BaFL pipeline does not include all of the possible ProbeSets, unlike those of RMA and dCHIP, and therefore care must be taken that the same genes are present in both the training and test data sets – in this case the number of input ProbeSets is given in the fourth column of Table 3.1.

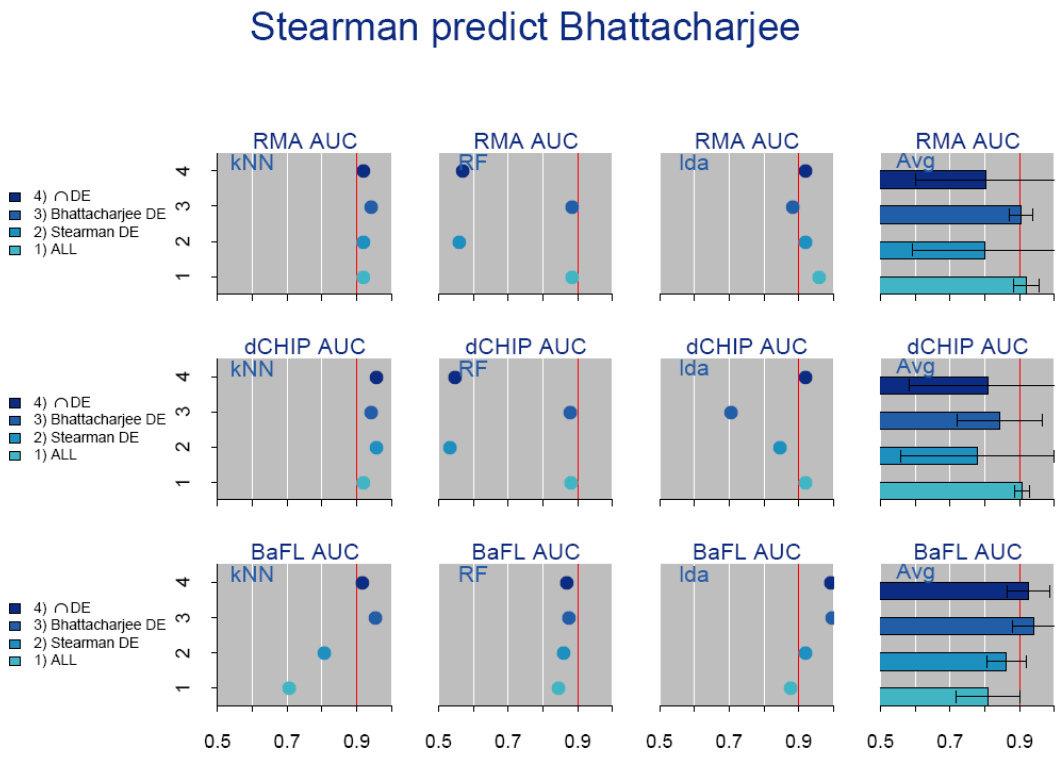


Figure 3.2: Down selection models- Stearman preditct Bhattacharjee. Classification results summarized, for kNN, RF and LDA classifier models, where training used the Stearman experiment-derived gene list as the initial input, and testing was done on the Bhattacharjee data. The last column summarizes the performance at each selection level across the three algorithms. Each graph shows the training set number (training sets described below) on the y-axis and the cumulative AUC value on the x-axis. Each row of graphs shows outcomes based on starting with a particular cleansing method (RMA, dCHIP and BAFL) followed by t-test classification for DE genes lists. The 4 sets of input genes result from additional selection criteria and are denoted by the different shades of blue circles in each figure. The 4 sets: set 1 was ALL, or no additional selection, the complete set of ProbeSets, 12,625 for the RMA and dCHIP algorithms and 4200 for BAFL data, set 2 was Stearman DE, the DE genes from the Stearman (RMA 5291, dCHIP 5208, BAFL 3344) experiment, set 3 was Bhattacharjee DE, the DE genes from the Bhattacharjee experiment (RMA 6595, dCHIP 6429, BAFL 480) and set 4 was the intersection of DE, the intersection of the DE genes in the two experiments (RMA 5085, dCHIP 4125, BAFL 325).

Figure 3.3 is the reverse experiment, using the cross-sections for the Bhattacharjee dataset-derived down selected ProbeSet lists and values generated by RMA, dCHIP and BaFL to train the models, with subsequent testing of the models' performance on the Stearman data.

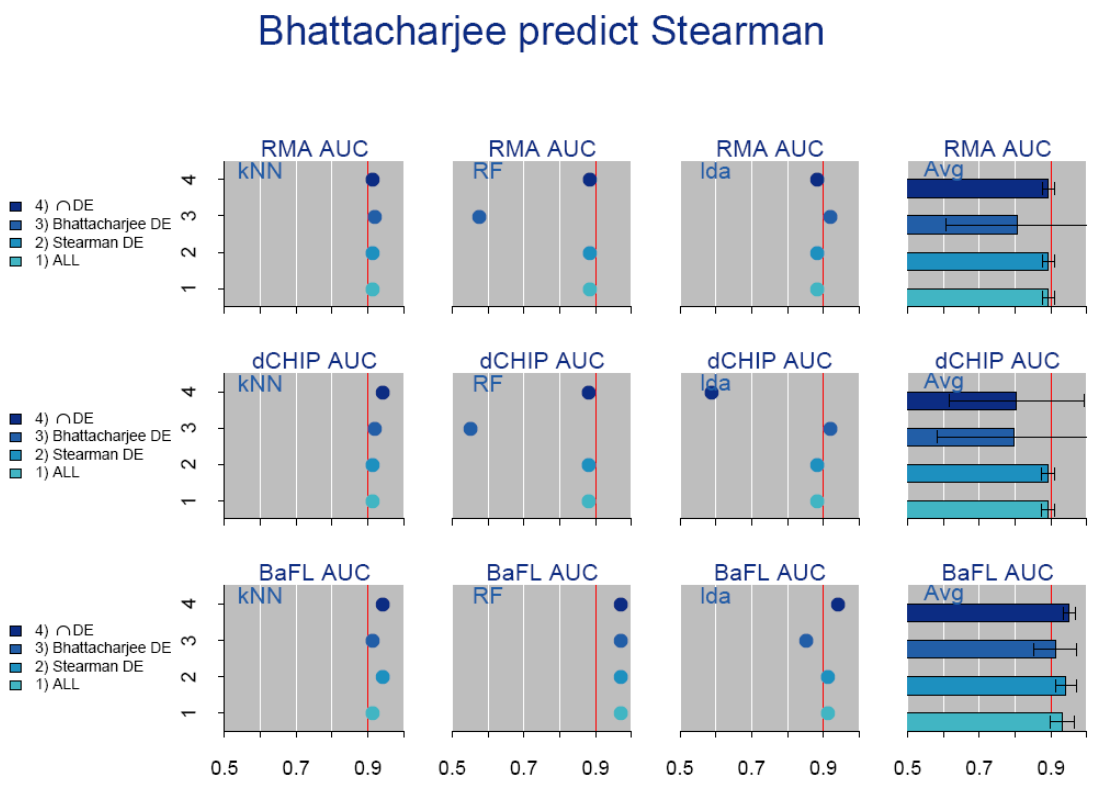


Figure 3.3: Down selection models- Bhattacharjee predict Stearman. Clustering results summarized, for kNN, RF and LDA classifier models, where training used the Bhattacharjee gene lists and testing was then done on the Stearman data. Each graph shows the training set number on the y-axis and the cumulative AUC value on the x-axis. The last column summarizes the performance at each selection level across the three algorithms. Each row of graphs shows outcomes for a cleansing method (RMA, dCHIP and BAFL) followed by t-test classification for DE results. The 4 sets of data resulting from additional selection criteria are denoted by the same colors and are the same as described in the legend to Figure 3.2.

Author's List (Validation)

The 325 DE ProbeSets were compared against the BaFL-passed genes in the author's published lists. The validation models incorporated minor perturbations, through random sampling of both the training and testing sets, over 100 iterations to achieve a reasonable measurement of the model's classification performance [6, 12, 36, 37]. Figure 3.4 presents the results of training the models using the Stearman dataset and testing on the Bhattacharjee data, and the three data

cleansing methods were used to generate the ProbeSet values. The same 3 classifications algorithms were employed as described above.

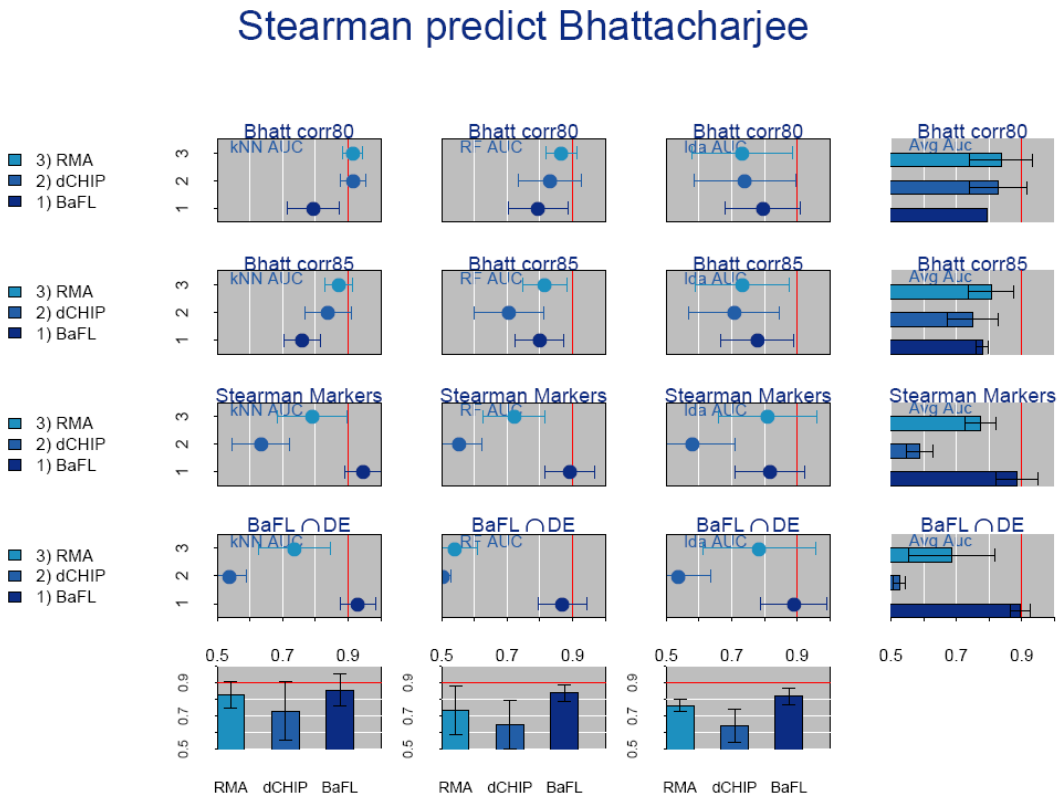


Figure 3.4: Candidate lists- Stearman predict Bhattacharjee. The genes from the original authors' papers proposed significant ProbeSet lists, further limited by those that met the BaFL cleansing criteria but not the Welch's t-test significance criterion, with ProbeSet values provided by one of the three data cleansing methods: these values were used as input to the three classifier models. Each graph shows the number of the ProbeSet value-generation method on the y-axis (described below) and the cumulative AUC value on the x-axis. Each row of graphs shows outcomes for a particular initial candidate ProbeSet list. Each column of graphs shows the results for a particular classifier, indicated in the graph. The red line highlights the 0.9 cumulative AUC. In this case the node indicates average AUC and the bars show the standard deviation for 100 random sampling iterations, with replacement. The final column summarizes each cleansing model's list interpretation, across the 3 classification algorithms. The final row of graphs summarizes the individual algorithms, per cleansing methodology, across all four 4 candidate lists and follows the same color scheme.

Figure 3.5 is the reverse experiment, where the gene list derived from the Bhattacharjee experiment was used for training the classification model, which was then tested on the Stearman

data. Classification is continued with the same models, using ProbeSet values from the probe cleansing methods, comparing the 325 DE BaFL ProbeSets to the authors' lists.

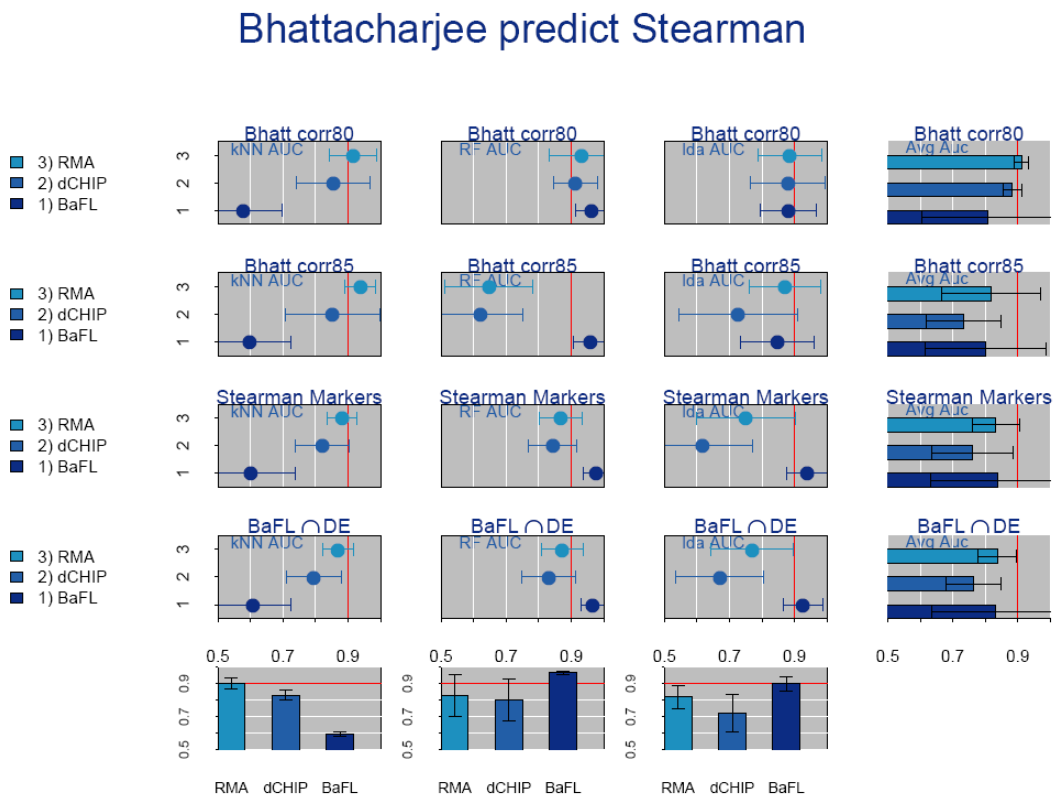


Figure 3.5: Candidate lists- Bhattacharjee predict Stearman. The genes from the original authors' papers proposed significant ProbeSet lists, further limited by those that met the BaFL cleansing criteria but not the Welch's t-test significance criterion, with ProbeSet values provided by one of the three data cleansing methods: these values were used as input to the three classifier models. Here training of the model used the (modified) Bhattacharjee ProbeSet list and testing was done using the Stearman data. Each graph shows the number of the ProbeSet value-generation method on the y-axis (described below) and the cumulative AUC value on the x-axis. Each row of graphs shows outcomes for a particular initial candidate ProbeSet list. Each column of graphs shows the results for a particular classifier, indicated in the graph. The red line highlights the 0.9 cumulative AUC. In this case the node indicates average AUC and the bars show the standard deviation for 100 random sampling with replacement repetitions. The final column summarizes each cleansing model's list interpretation, across the 3 classification algorithms. The final row of graphs summarizes the individual algorithms, per cleansing methodology, across all four 4 candidate lists and follows the same color scheme.

Discussion

A striking result from this series of experiments is the significant improvement in the power of the t-test for the Bhattacharjee dataset, when the ProbeSets to be considered and the value of those ProbeSets are produced using the BaFL pipeline [39, 40]. This increase in power was not observed for any of the cleansing methodologies with the smaller Stearman dataset, which, is much smaller, although it is completely replicated. From Figure 3.1 it can be seen that there is a significant improvement in the uniformity of the *p* value kernel distributions of the BaFL-generated ProbeSet values tested for significant differential expression for the Bhattacharjee dataset, compared to those of RMA or dCHIP, particularly for the null *p* values [39]. Since both datasets show similar variance after BaFL processing (see Chapter 2), and the Bhattacharjee dataset is both more heterogeneous (disease stage) and less precise (less replication) than the Stearman dataset, the lack of power comes down to the difference in sample size of the experiments [5].

Models and Class Predictions

The impact of the lack of power becomes apparent in the performance of the down-selected RMA and dCHIP-based classification models, where the resulting datasets have little improvement or even a loss of performance when the t-test down-selection is used (compare the ALL set to the other three Sets in the top two rows of graphs in Figures 3.2 and 3.3). A meaningful outcome would show a gain in information when going from 12,000 to ~6,000 genes, where the method has allowed the genes with an impact on the phenotype to be retrieved [6, 41]. This did not occur when starting with the RMA and dCHIP cleansing methods, in spite of trying three types of classification models in the search. In fact, there is a consistent increase in the variation across the

classification models coinciding with down selection, as demonstrated in the last column. This phenomenon mirrors what has been observed generally with Microarray data, the poor prediction performance of proposed gene lists given different data and classification approaches [37, 42, 43]. Gains are much more consistent, if not large, when the BaFL-cleansed t-test down-selected data are used (the third row of graphs in Figures 3.2 and 3.3).

Additionally, of the three models employed, Random Forest consistently did a poor job for the RMA and dCHIP ProbeSets which were generated as the DE ProbeSets for the training model; however, this was not observed with BaFL-generated values. This suggests that during the stochastic development of the decision trees the selection of important features is often specific to the dataset and not the disease condition. When the ProbeSet response is variable, its importance to different models can either diminish, weakening its role as a classifier, or the regulation pattern can be inverted, generating conflicting classifications. The RMA and dCHIP interpretations of the datasets present 1.8% and 3.6% (respectively) of the intersecting DE ProbeSets with conflicting regulation patterns between the two datasets, as shown in red in Figure 3.6. Linear discriminant analysis, which weights all the features and hence allows variable genes to be more or less important, did not show the same sensitivity as Random Forest.

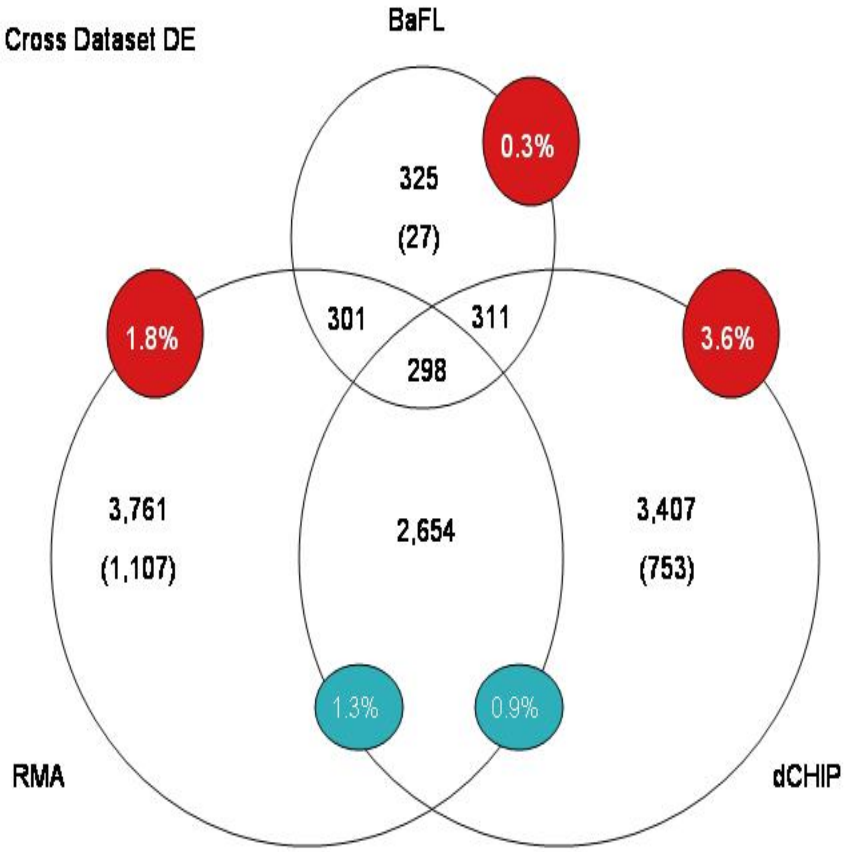


Figure 3.6: Concordance summary. The number of DE members of the ProbeSet lists that are shared in the designated categories between the Bhattacharjee and Stearmen experiment results. Numbers in non-overlapping regions are the total list sizes and the numbers in parentheses below are the number of ProbeSets unique to that category. Numbers in overlap regions reflect shared ProbeSet list members. The red-circled numbers show the percentage of these DE ProbeSets which have contradictory regulation patterns ($\mu 1 > \mu 2$) between datasets for each cleansing methodology. The light blue circle shows the same for the intersecting DE ProbeSets.

While performance improvements are not always larger than that obtained with the complete data set (except for the kNN model), for all three classification approaches we demonstrate some improvement using the BaFL-pipeline generated values and any of the candidate gene lists. We do see a large performance gain associated with the removal of uninformative ProbeSets for the LDA model. The fact that there is no loss, and in some cases a gain, in performance with a ten-

fold decrease in the number of candidate genes is important for diagnostic applications [6, 41]. It is also useful that there is no great dependence on a particular model for doing the classification. Most notably, we see very good cross-experiment performance, even with quite limited candidate gene lists, a rare achievement with Microarray data.

Author's List (Validation)

Of the 72 AUC scores recorded for classification performance, the values generated by the BaFL pipeline when used for candidate gene lists achieved 5 of 6 best overall scores (87-97%) over the three classification algorithms, although not all of these are significant given the variance in the results. The 6th case, and the exception, was the kNN results when Bhattacharjee DE candidate genes, using the RMA generated values, were used to predict Stearman data (93.88%). In addition the BaFL-defined DE ProbeSets achieve the highest AUC for almost every individual analysis of the BaFL intersecting DE and Stearman list, again the sole exception is the kNN implementation of Bhattacharjee predicting Stearman. If you average the performance across the lists per cleansing routine, values obtained using the BaFL method achieved the highest performance 5 of 6 times (bottom row of Figures 3.4 and 3.5). Equal or improved performance across experiments, with less variability and smaller candidate gene lists, and low sensitivity to the model, are diagnostic goals that the BaFL method appears to be well positioned to achieve. We note that a possible cause for the relatively poorer BaFL performance for Bhattacharjee predict Stearman implementation of kNN, may be the result of random replicate removal out of the small Stearman dataset, coupled with the absence of scaling across arrays in the BaFL pipeline, as compared to RMA and dCHIP pipelines, which do incorporate scaling steps. We support this statement by noting that the degradation of performance was not observed in the

whole model analysis without permutation, where BaFL achieved an average AUC of 91.2% ± 3.4% across the 4 lists.

The analysis of the results we obtained using subsets of the author's candidate lists leads to a number of observations. First, the gene list based on the more stringent correlation coefficient in the Bhattacharjee experiment (0.85 corr) consistently underperforms with respect to the results for the less stringent correlation coefficient. These two Bhattacharjee lists share the fewest common DE classified genes between datasets. According to the original article, these ProbeSets were selected as having the highest Pearson correlation values with respect to phenotype across 45 adenocarcinoma sample replicates [3]. This should result in more consistent expression value changes, but as summarized in Figure 3.7, there is more disparity in the cross-experiment fold changes calculated for RMA and dCHIP including the ProbeSets which are not differentially expressed. Figure 3.7 presents the comparison of fold changes (disease/normal) of each ProbeSet identified by the 4 filtered candidate lists, with the DE ProbeSets identified by arrows. The columns are the 4 candidate lists: Bhattacharjee Correlation80, Bhattacharjee Correlation85, Stearman markers, and our 325 DE ProbeSets, respectively. The rows are the three cleansing approaches: RMA, dCHIP, and BaFL, respectively. Uninformative ProbeSets appear as grey circles near 1 and linear regression of the data using QR decomposition [16] is drawn for orientation.

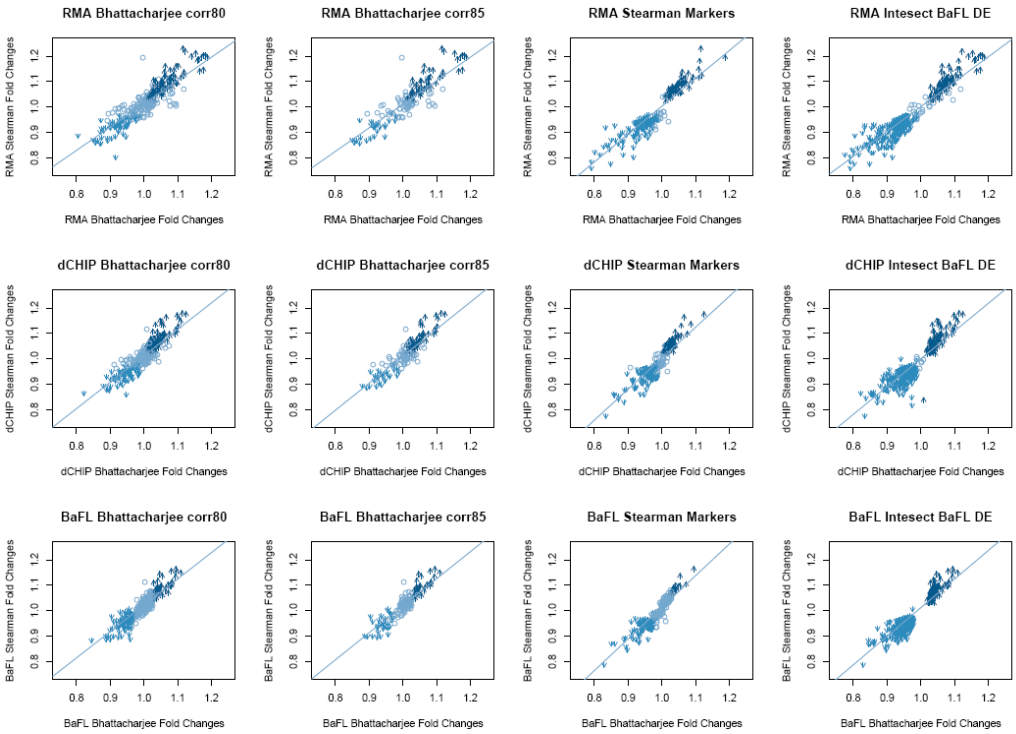


Figure 3.7: Candidate lists fold change differences. Comparing ProbeSet fold changes for candidate list ProbeSets between the two experiments. For each class of samples in each experiment, ProbeSet values are calculated by one of the three data-cleansing methods and the mean is found, after which the Normal:Disease class fold change is calculated. The y-axes are the mean fold-change for the Stearman experiment and the x-axes are the mean fold-changes for the Bhattacharjee experiment. The line gives the QR regression slope for the points. A perfect 45 degree slope would indicate that the fold-changes are on average the same across the experiments, although individual ProbeSets might change, while a shift in the slope means that there is a consistent bias in the direction that fold changes are seen for one experiment relative to the other, given a particular method for generating values. Those ProbeSets that show a negative fold-change in both experiments are shown with a medium blue down arrow, those for which both show a positive fold change are shown with a dark blue up arrow, and those that have differences in the direction of fold change are shown as grey circles. Also observable is an exaggerated funneling patterns for DE ProbeSets in the RMA and dCHIP interpretations of Stearman markers and the Intersecting BaFL DE ProbeSets, intimating that the more significant the expression pattern the more divergent the actual level of fold-change becomes across datasets.

Another curious feature is a funnel shape that is most evident in the top graph (RMA). The effect is more pronounced in the larger, down-regulated group of ProbeSets and reflects the fact that the larger the change the more likely the two experiments are to disagree as to the extent of that change, but that the effect is symmetrical: one set is not more likely to consistently have a larger

change, apparently the greater variance in the Bhattacharjee data is fairly randomly distributed in these sets of genes. In the fourth column the BaFL intersection with the other three candidate gene lists is used as the new candidate gene list. There is a decreasing number of uninformative ProbeSets from RMA to dCHIP to BaFL, which of course has none. The symmetrical funnel shapes of the two clusters of ProbeSets with up- or down-regulation are apparent in all three graphs, some of which can be attributed to biological variation. However, this variability appears to be least significant for the BaFL interpretation.

The fact that a large fraction of the genes in the Bhattacharjee candidate gene lists are not differentially expressed derives from the candidate selection approach employed [3]. The selection of consistently expressed ProbeSets for the adenocarcinoma phenotype would not imply that the every ProbeSet is a strong classifier and the only clear trend for these lists is that the larger, less significance-selected list performs better. For a candidate list defined in this way no cleansing methodology appears to yield a better outcome (Figures 3.4 & 3.5). In contrast, the Stearman markers have a markedly larger group of DE ProbeSets, which most likely reflects the fact that the significance of those genes was assigned for two reasons: both differential expression (in a single experiment) and comparative genomics (correlation to a mouse model) [4]. Table 3.2 gives the total ProbeSets and the fraction in the two DE categories for each of the candidate lists and ProbeSet value generation methods. It is clear that the greatest concordance comes about when members of gene lists are selected for meeting criteria across multiple experiments (both Stearman and BaFL used two, albeit in different ways). When the Stearman marker subset was used as the candidate gene list the RMA and dCHIP-generated values led to better overall concordance in the DE prediction (Table 3.2 row 4) which may reflect expression characteristics of this set of genes (reliable and stable expression) [4]. For these doubly-selected gene lists we

do observe that the BaFL interpretation outperforms dCHIP and RMA, particularly if centering and scaling the BaFL data further improves the classification accuracy, as we expect.

Table 3.2: Candidate list concordance numbers. Numbers of ProbeSets in candidate gene lists, starting with the original list (column 2) with the filter of passing the BaFL filtering criteria (column3), and the percentage of those (column 3) ProbeSets which show concordant differential expression classification between datasets per cleansing methodology (last three columns).

	List Total	BaFL retention	RMA % DE Concordance	dCHIP % DE Concordance	BaFL % DE Concordance
Bhatt 80%	675	267	54.68%	56.55%	31.46%
Bhatt 85%	366	136	58.09%	59.56%	36.76%
Stearman	409	178	92.13%	80.90%	44.94%
BaFL ∩ DE	325	325	92.62%	95.69%	100.00%

Conclusion

Although clustering and weighting schemes differ, in general the information content of a gene with respect to phenotype is believed to be larger when it is significantly differentially expressed and always in the same direction (i.e. up or down) across experiments. This many not be true for multigenic interactions, or where an effect is due to a larger range of expression rather than a particular level or direction of change, but most methods rely on the first assumption. It is true that genes that are expressed but do not change are not at all informative. If the measure of importance is linked to differential expression then the method that most consistently shows the same genes changing under the experimental conditions, in the same direction and to the same extent, will be a preferred method. The BaFL pipeline values have been shown in this chapter to deliver this result. We note here that in this chapter we used the aggregated ProbeSet values, even

though in Chapter 2 we showed that these disguise what are really two classes of change, S and DE. This was done because there is no way to replicate that level of discrimination with the RMA and dCHIP pipelines, and the goal in this chapter was to compare the effects of using the pipelines. We suggest that it is *how* the data is integrated that is causing some of the reproducibility issues with Microarray data, not the data itself, at least for within-platform reproducibility. As shown in Figures 3.4-5, there is an observable decrease in classification performance that occurs when using the more rigorously pruned (and shorter) gene list, which is counterintuitive if the greater rigor really led to greater quality. It seems likely that the greater rigor is really selecting for lab-specific and experiment-specific factors, rather than sample state relevant factors. Demonstrating that the BaFL pipeline is a more effective cleansing approach than RMA and dCHIP is a difficult task when the underlying cause of differences cannot be exactly isolated, and when the sources of data are clinical samples with incomplete and variable levels of replication. Direct comparisons recapitulate a result seen by others, namely that RMA and dCHIP are not consistent with one another within a dataset, nor able to perform well across experiments, while BaFL shares about the same overlap with each. Nevertheless, in this chapter we have provided evidence that BaFL-pipeline ProbeSet values followed by a simple t-test for differential expression yields a more effective candidate gene list for training a model for classifying the results of additional experiments that do the competing methods. The advantages for disease diagnostic purposes are the much smaller size of the candidate gene list and the relative lack of sensitivity to the specific type of model. As with any clinical experiment a larger sample size leads to more robust results, and a meta-experiment, with at least two independent experiments, is most effective. It was also of note that the most stable differentially expressed genes are not necessarily the most informative for the disease phenotype.

Chapter 4: Data Mining

Adenocarcinoma is a non small cell (NSCLC) lung cancer sub-type, and the most frequent type of lung cancer found in the world today [1, 2]. Adenocarcinomas are peripherally located in the lungs and develop from clara cells, alveoli, and mucin producing cells [1]. While, tobacco smoking has been well established as an initiating condition for lung cancer, with 80-90% of lung cancer cases arising in tobacco smokers, adenocarcinoma in particular is most common among women, non-smokers, and the young [1]. Given that the incidence rate of adenocarcinoma is increasing and affecting non-traditional patients, understanding the disease is of immediate concern [1, 2].

Using the methods described in the previous chapters, we have created two 2-class datasets, one from a subset of the Bhattacharjee dataset and the other from the original Stearman (human subset) experiments [3, 4]. In this chapter we begin with the down-selected ProbeSet list, presented in Chapter 3, consisting of those 325 differentially expressed ProbeSets common to both datasets. The values that the BaFL pipeline yields for these ProbeSets lead to datasets with considerable latent structure; we will demonstrate that this latent structure is superior to that of RMA and dCHIP supplied values using two widely-accepted dimensionality reduction methods: Principal Components Analysis [5, 6], which is linear , and a Laplacian method which is non-linear [7]. In validating the results of these analyses, we use sample correlation to explore the gene/ProbeSet clusters.

Although a list of 325 candidate genes is not large by the standards of Microarray experiments, and may reflect real biological contributions to a complex phenotype, to produce a practically useful diagnostic test one would like the smallest possible list of genes that have strong effects. There are many ways to additionally down-select features, one of which was used by the investigators of the Stearman experiment, comparative genomics [4]. Here we chose to use a more traditional statistical approach, in which feature selection from the 325 ProbeSets was accomplished by incorporating the Bonferroni correction [8, 9]. The Bonferroni is a stringent correction to accommodate the multiple hypothesis tests. The underlying assumption of the Bonferroni correction is that all null hypotheses are true (the mean expressions are equal) [8, 9]. Thereby, only the ProbeSets with extreme differences in the mean ProbeSet intensity will survive the correction. There has been considerable debate as to whether the extreme rigor of the Bonferroni method is appropriate for expression Microarray experiments, and one of the questions is how much potentially valuable information we lose by applying this technique [8]. As before the performance of the candidate gene list in classifying samples is assessed using kNN, LDA and RF classifiers [10-13] to build models that are then judged by AUC scores [14, 15].

Materials and Methods

As previously described, the 'Bhattacharjee', dataset [3], consists of 17 normal and 237 diseased samples, including 51 adenocarcinoma replicates, with disease category assigned after histopathological examination. From this study we used 125 of the 190 adenocarcinoma array results and 13 of the 17 normal results; the selection criteria are described below. The second, 'Stearman', dataset (http:/www.ncbi.nlm.nih.gov/geo/; accession number GSE2514) consists of 39 tissue samples, all replicated, from 5 male and 5 female patients (four samples were taken

from each patient: 2 normal looking that are adjacent to the tumor and 2 adenocarcinoma samples); one of the normal samples is missing, presumably for high tumor content [4]. These sample biopsies were harvested using microdissection techniques and then snap-frozen [4]. The final BaFL interpretation of the sample included 17 tumor samples and 14 normal (i.e. adjacent) samples. These two datasets had a common group of 325 ProbeSets that were identified as significantly differentially expressed between the diseased and normal states, based upon Welch's t-test of BaFL-provided values, applied at the ProbeSet (aggregate) level, described in detail in Chapter 3 [9]. The latent structure of these 325 ProbeSets is presented as a final piece of evidence showing why these ProbeSets should serve as the foundation for the feature selection, and particularly why the BaFL interpretation of these ProbeSets' signal intensities is preferred.

Feature Selection

Using the BaFL-provided values for these genes, the Bonferroni correction was applied to each dataset; correction cutoffs were less than 1.18e-05 for the Bhattacharjee data and 7.68e-06 for the Stearman data.

$$Q = p / t \,, \tag{4.1}$$

where $\{Q, 1\}$ if p is less than the correction. The number of total tests is represented by t, and p represents the p-value of Welch's t-test [8, 9]. This resulted in 79 Bonferroni corrected ProbeSets out of the 480 Bhattacharjee DE ProbeSets and 352 Bonferroni corrected ProbeSets from the 3,162 Stearman DE ProbeSets; the intersection set was 30 ProbeSets. The Bonferroni correction was applied also to both original BaFL-interpreted 4,253 Bhattacharjee ProbeSets and 6,506 BaFL interpreted Stearman ProbeSets. The resulting intersection of 'most significant' ProbeSets

was determined to consist of 30 ProbeSets, and represents a subset of the original 325 differentially expressed ProbeSets. An additional ProbeSet, 34342_s_at (SPP1), was added because it was determined to be 'on' in the Bhattacharjee adenocarcinoma samples and 'off' for the Bhattacharjee normal patients, resulting in a final candidate list of 31 ProbeSets.

Survival Curves

Survival curves for the osteopontin (SPP1) gene used each of the three value-generating methods (log-transformed the BaFL values), for the Bhattacharjee experimental data, were generated with the R `maxstat` package [16, 17]. The SPP1 ProbeSet intensities were associated with the supplementary clinical data [3], and are provided in the Supplementary Materials Data folder. Survival curves plot the survival probability of a patient within a cohort over time [18].

Classifiers and AUC Performance Metrics

The AUC classification performance of these 31 ProbeSets was assessed exactly as described in Chapter 3, using the linear discriminant analysis (LDA), Random Forest (RF), and k Nearest Neighbors (kNN) classification algorithms [10-13]. These classification experiments used random sampling with replacement of both the training and test datasets for 100 permutations [10]. Having selected these ProbeSets as described above, the values obtained from RMA, dCHIP and BaFL were used as input to the classifiers and in the results below the outcomes are compared to those obtained in Chapter 3.

Results

The results sections is divided into thirds, with the latent structure presented first as final supporting evidence as to the quality of the BaFL-DE-intersection 325 ProbeSets. The next section presents the analysis supporting the inclusion of SPP1, osteopontin, in the candidate list; briefly, the ProbeSet was identified by the BaFL pipeline as being 'on' for the adenocarcinoma and 'off' for the normal samples. Only two such on/off genes were identified in these data sets. The final section shows the performance results of the candidate gene lists in independent validation across the two datasets, where the BaFL interpretation is benchmarked against the RMA and dCHIP interpretation for the same list of 31 candidate ProbeSets.

Latent Structure

Principal component analysis (PCA) of the latent structure of 325 candidate genes was performed to project the genes into 2 dimensional space, based upon the correlation between samples. The RMA and dCHIP algorithms center and scale their data to better approximate a normal distribution, while the BaFL produced data is log transformed to achieve a similar approximation. For all analysis the data was centered through the R functional parameter (center=T), although scaling was not done (scale =F). Presented in Figure 4.1 is the PCA latent structure for the Bhattacharjee dataset for the three sets of values.

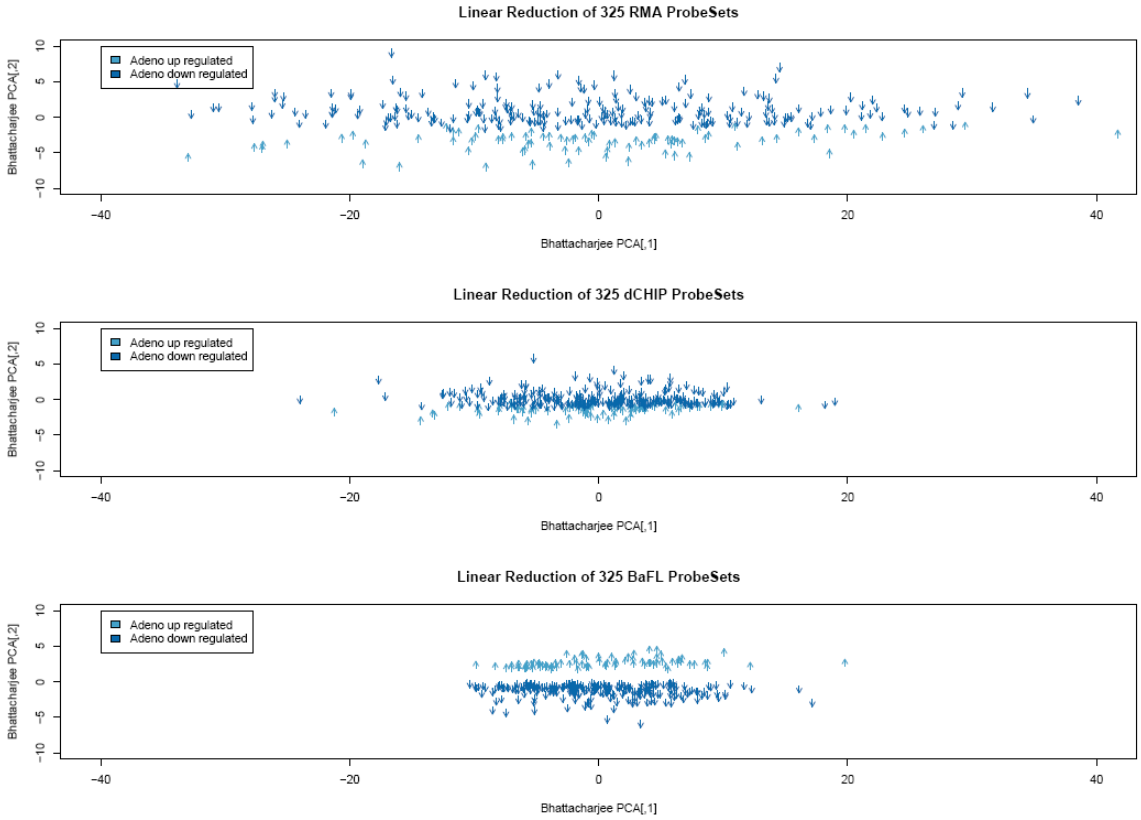


Figure 4.1: Non-traditional PCA analysis of Bhattacharjee data. PCA analysis of the Bhattacharjee data for ProbeSets based upon sample correlation, using singular value decomposition of the data matrix (R prcomp function) [17]. Latent structure can be observed in all three graphs. Top to bottom: RMA, dCHIP, and BaFL produced values used as input to the data matrix.

PCA is a linear method, but not all gene relationships are linear [5-7]. The results of using a Laplacian, non-linear reduction, method are presented for the Bhattacharjee data (Figure 4.2), applied according to [7]. In this approach single value decomposition (SVD) of the correlation matrix is performed [17] and normalized to the first column (ProbeSet) [7]. The first 2 Laplacian dimensions are then projected into 2 dimensional space. Again, the BaFL produced data was transformed with the simple log first.

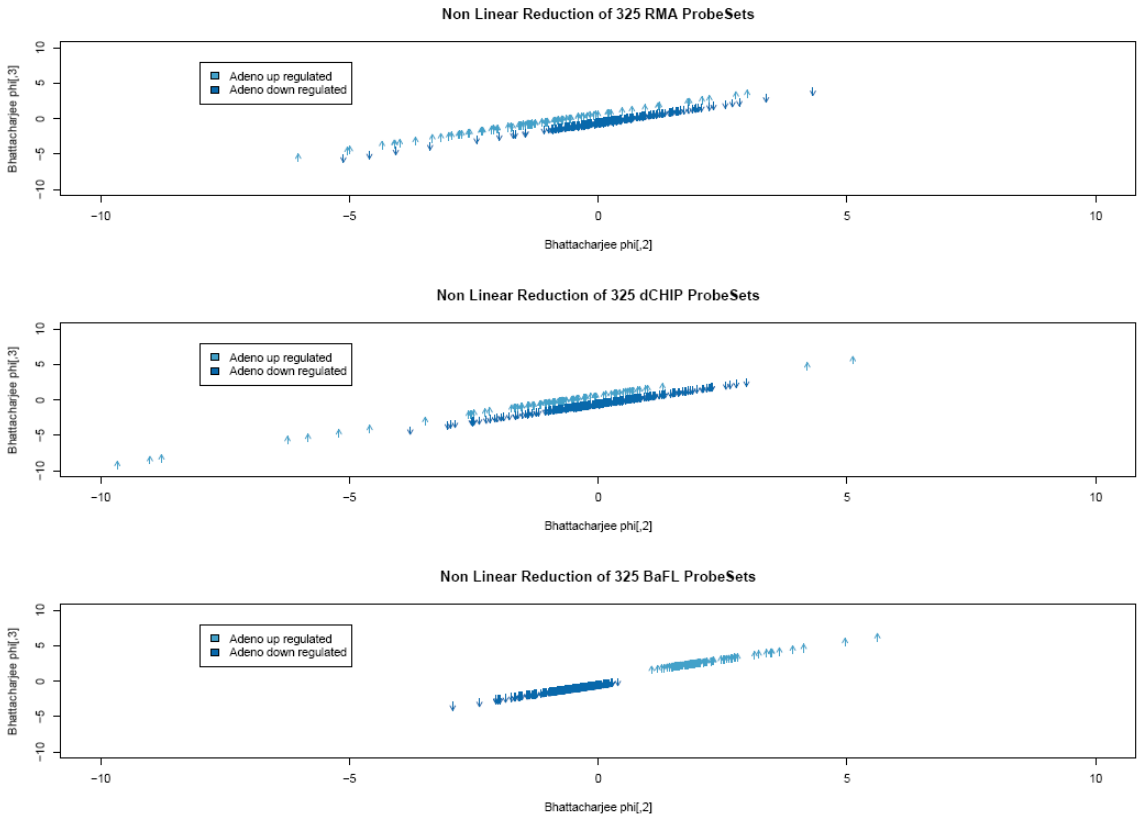


Figure 4.2: Non-traditional Laplacian dimension reduction. Results of the Laplacian dimension reduction method for the Bhattacharjee data, showing ProbeSet correlation by sample class. The direction of change of the ProbeSet is indicated both by color and arrow. Latent structure is observed in all three groups. Top to bottom the graphs are from data matrices that started with values derived from the RMA, dCHIP, and log-transformed BaFL values.

The structure observed via the Laplacian dimension reduction method was assessed across both datasets to see if the structure was consistent in both. There was significant preservation of the structure in the case of the BaFL interpreted data, but in the RMA and dCHIP data sets the structure was lost, as presented in Figure 4.3. Statistical discordance indicates those ProbeSets which were found to be significantly expressed in only one of the datasets and applies only to the RMA and dCHIP produced values for these 325 ProbeSets. Pattern discordance, on the other hand, indicates ProbeSets which demonstrated differential expression in both dataset, but the direction of change was different per dataset. While there does appear to be some latent structure

for the RMA and dCHIP interpretations, the separation is not complete as it is for the BaFL results. Since, the latent structure that is presented by the BaFL interpretation could be an artifact of normalizing the SVD matrix to the first ProbeSet, permutations were performed upon the ProbeSet ordering. Permuting the gene order did not affect the latent structure that is observed, although minor variations of the structure were found. These results are presented in the Appendix E.

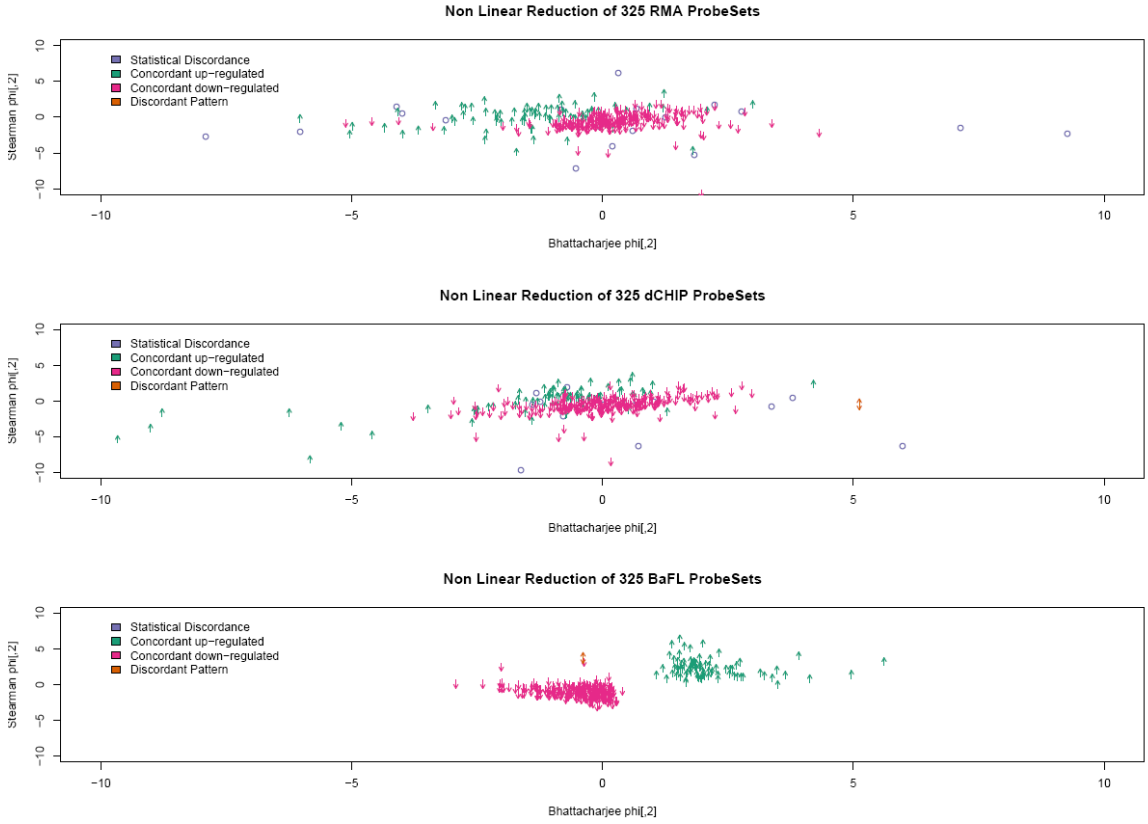


Figure 4.3: Cross dataset latent Laplacian structure. Results of the Laplacian dimension reduction method on both the Bhattacharjee and Stearman data for ProbeSets, based upon sample correlations. These graphs compare the 1st Laplacian dimension of both datasets, Stearman on the y axis and Bhattacharjee on the x axis. The direction of change is indicated by the color and direction of the arrow; there are four colors to account for incongruencies in the direction of change between the datasets, as noted in the text. Latent Structure is apparent in the third graph of the BaFL interpreted data. From top to bottom: RMA, dCHIP, and BaFL produced values were used in the original data matrices.

In the above ProbeSet group it is clear that the structure depends greatly on the associated value, since for a large majority of the ProbeSets the *direction* of change is the same regardless of the method for generating the value. However, the selection of the ProbeSets was based on BaFL values: it is possible that an intersection of RMA/dCHIP DE ProbeSets would have equally meaningful structure. To assess this, a list of predicted DE ProbeSets shared by RMA and dCHIP in both experiments was derived. As a final step, that list was filtered for those retained (not necessarily DE) in the BaFL protocol. The resulting list has 940 ProbeSets. The Laplacian dimension reduction method was again employed using values supplied by each of the three methods; results are presented in the Appendix F. Similarly these 940 ProbeSets demonstrated the same consistent latent structure for the log transformed data, while The RMA and dCHIP produced data lacked any structure. Additionally, these 940 ProbeSets possessed a marked increase in the number of pattern discondant ProbeSets for the RMA and dCHIP data.

We note that for one ProbeSet the BaFL interpretation of the Bhattacharjee and Stearman results disagreed as to the direction of differential expression. This ProbeSet, 34532_at / CUGBP2, is the overall pattern is reflected across all of the individual probes as well, and is not the result of a single highly variant probe in the set (data not shown).

The 'On' and 'Off' Category of ProbeSets

Because the BaFL-based pipeline requires that a minimum of 4 probes be present before a ProbeSet is considered it will miss any genes that are always present in one sample class and never present in the other. For the goal of finding the smallest number of ProbeSets with the greatest difference between classes (the ideal diagnostic) this is a flaw, if such ProbeSets exist. In fact, we have identified only two such Probesets in these experiments. Osteopontin (34342_s_at)

was one of them: it is counted as present in the adenocarcinoma samples and not detectable in the normal samples. The source of the discrepancy is the linear range filter: the 7 probes with reliable values in the diseased sample fall below that threshold in the normal samples. This ProbeSet alone can classify the Bhattacharjeedata with a 98.9% AUC, as even if you allow aggregation of probe values below 200 fluorescent units, only three of the normal samples have ProbeSet values greater than 200 f.u. Figure 4.4 presents boxplots of the BaFL-based ProbeSet means of the osteopontin log-transformed expression values for each of the five tissue classes in the Bhattacharee experiment. This figure mirrors that of Hu's 2005 plasma protein study on lung cancer [19]. The aggregated ProbeSet values used only the 7 probes which survived the cleansing for the adenocarcinoma samples, although it should be noted that in the squamous carcinoma and small cell cancer samples additional probes were retained.

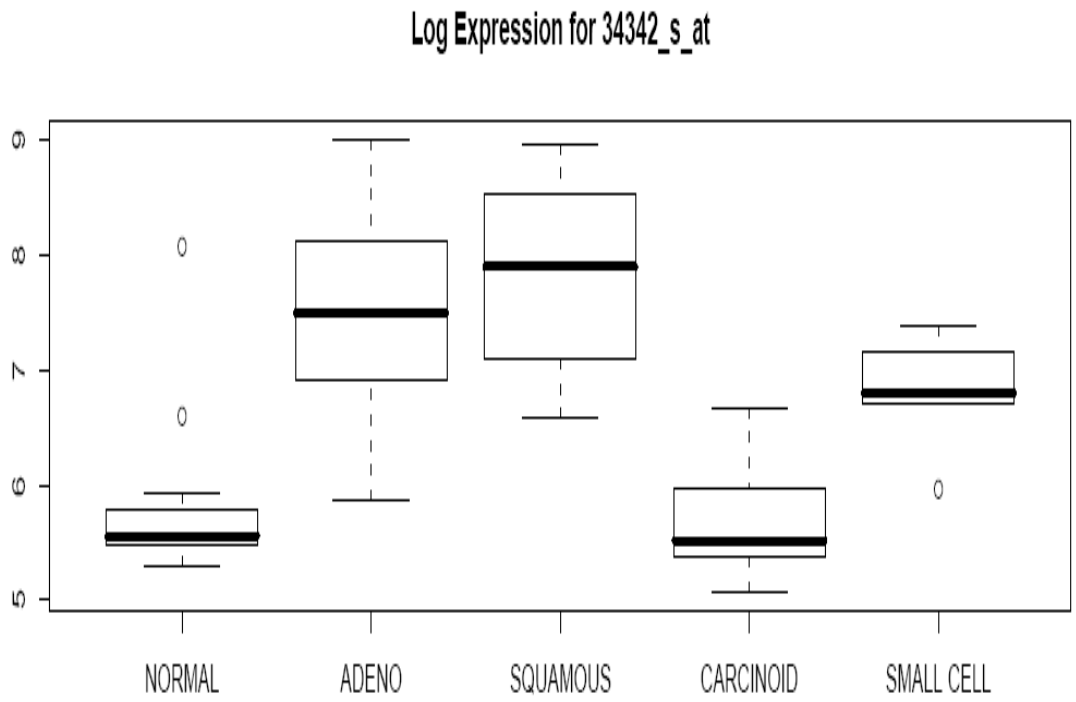


Figure 4.4: Signal intensity boxplots for Osteopontin. Boxplots of the log expression distributions, based on aggregation of BaFL values, for ProbeSet 34342_s_at (SPP1). The Bhattacharjee data are shown. The y-axis is the log-transformed BaFL value (natural log) of the expression values of all samples in the class, and the x-axis indicates the sample class (the number of samples is very small in all class but the adnocarcinoma). This figure is very similar to results shown in the 2005 plasma protein level study of osteopontin in various lung cancers [19].

The similarity to the plasma protein levels reported by Hu, *et al.* is not matched by similarity to the survival rates, or significance with regard to stage progression, reported by Donati, *et al.*, shown in Figure 7 [20]. The R `maxstat` package was used to generate the following survival rate figures [16].

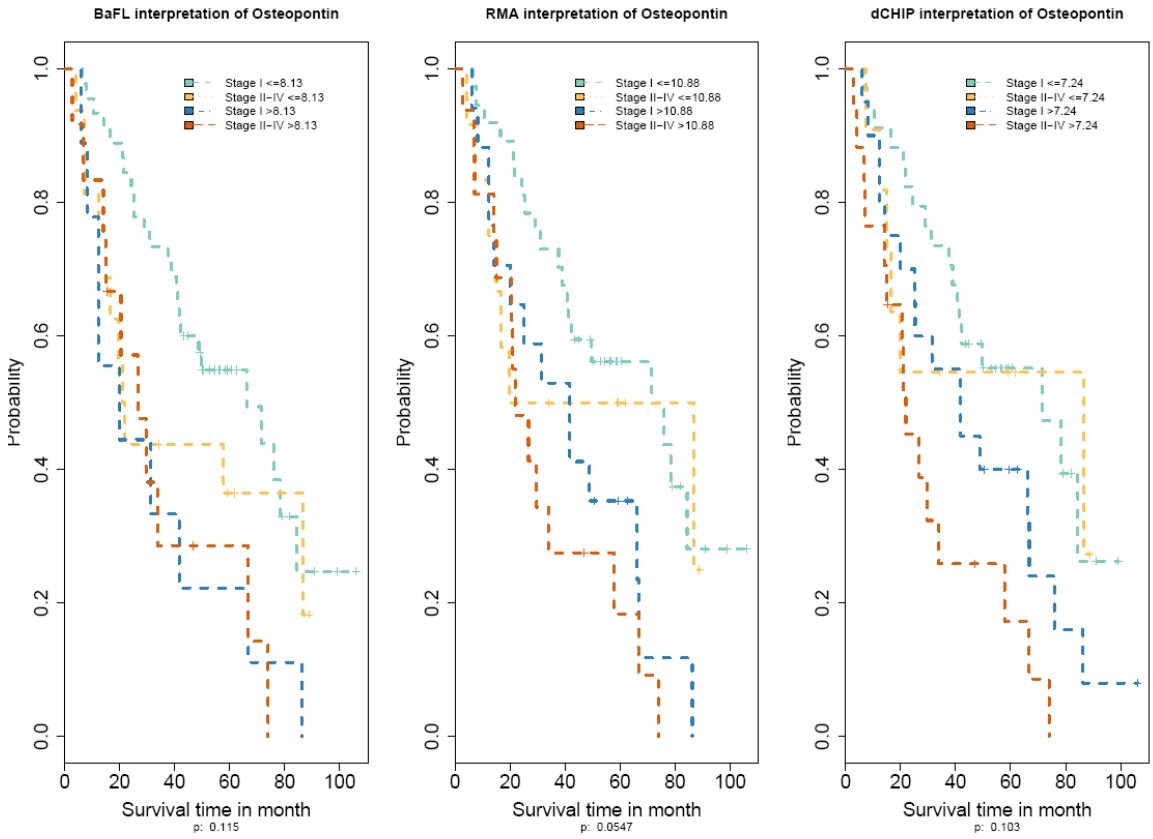


Figure 4.5: Kaplan-Meyer survival curves for Osteopontin. ProbeSet 34342_s_at (SPP1) survival rates with respect to stage progression. The significance of the osteopontin expression levels is presented in the sub label along the x-axis: RMA came the closest to a p-value of 0.05. The cutoffs in the legends represent the log expression threshold which divides the sub-populations. Note that in the far right graph dCHIP signal interpretation of the data an individual stage I patient has switched patient groups, having an expression level above the 7.24 cut-off.

After examining the differences in survival statistics and taking into account the relatively elevated expression level of osteopontin in the three previously mentioned normal samples, we further divided the data. Support for this decision can be found in the maximum tumor pathology as reported by Bhattacharjee, *et al*. in the supplementary data to that article [3]. A mean cut-off of 75% tumor pathology was established; the consequent associations between osteopontin expression level, tumor pathology, and stage are presented in Figures 4.6 and 4.7. Figure 4.6 is the level of osteopontin expression compared to patient survival for patients with greater than

75% tumor pathology for both stages I and II-IV, again using all three methods for generating the expression level.

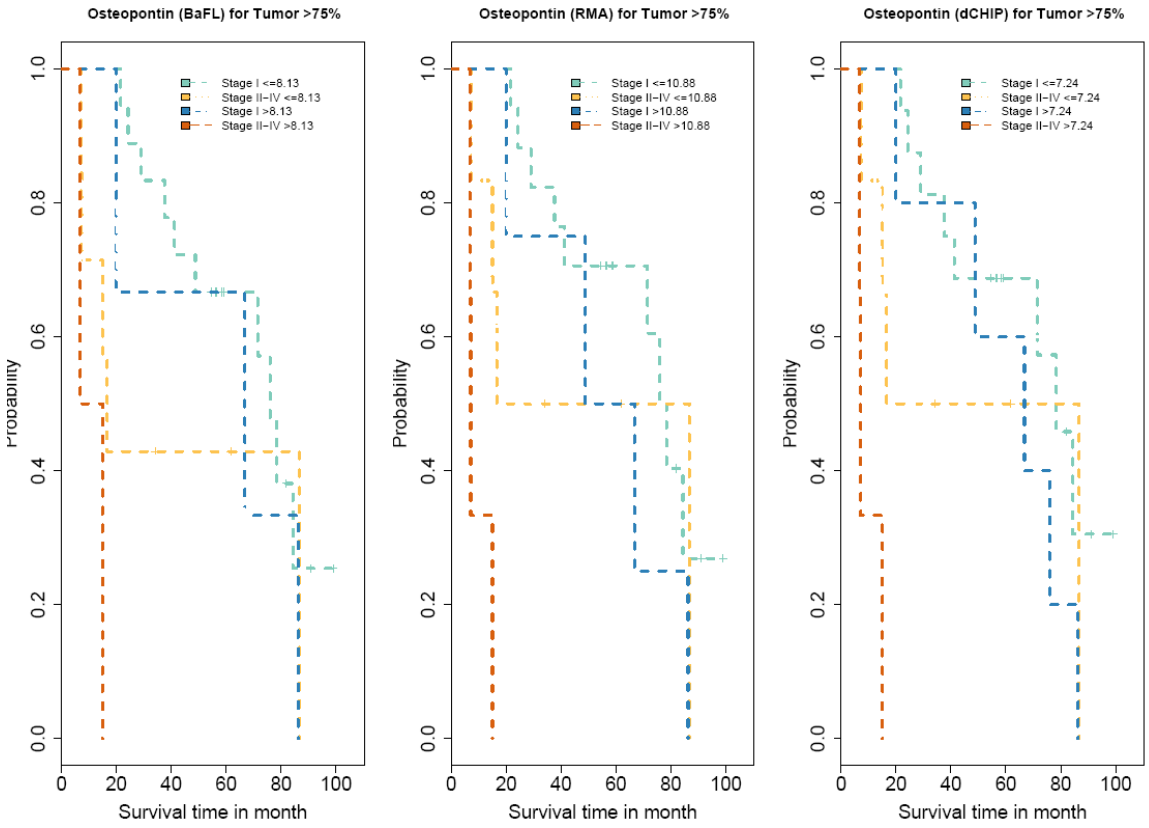


Figure 4.6: High grade tumor survival rates. ProbeSet 34342_s_at (SPP1) and patient survival rates with respect to greater than 75% tumor pathology and stage progression. All three probe cleansing methodologies show that the level of osteopontin has significant correlation with survival across all stages.

Figure 4.7 shows the outcomes for the other set of patients, those with a low grade (less than or equal to 75%) tumor pathology and the impact of elevated osteopontin and the progression of cancer. A dramatic effect is observed for stage I cancer patients with high osteopontin levels, with levels provided by the BaFL methodology. Neither RMA or dCHIP provided levels show this important linkage between cancer progression and elevated osteopontin levels.

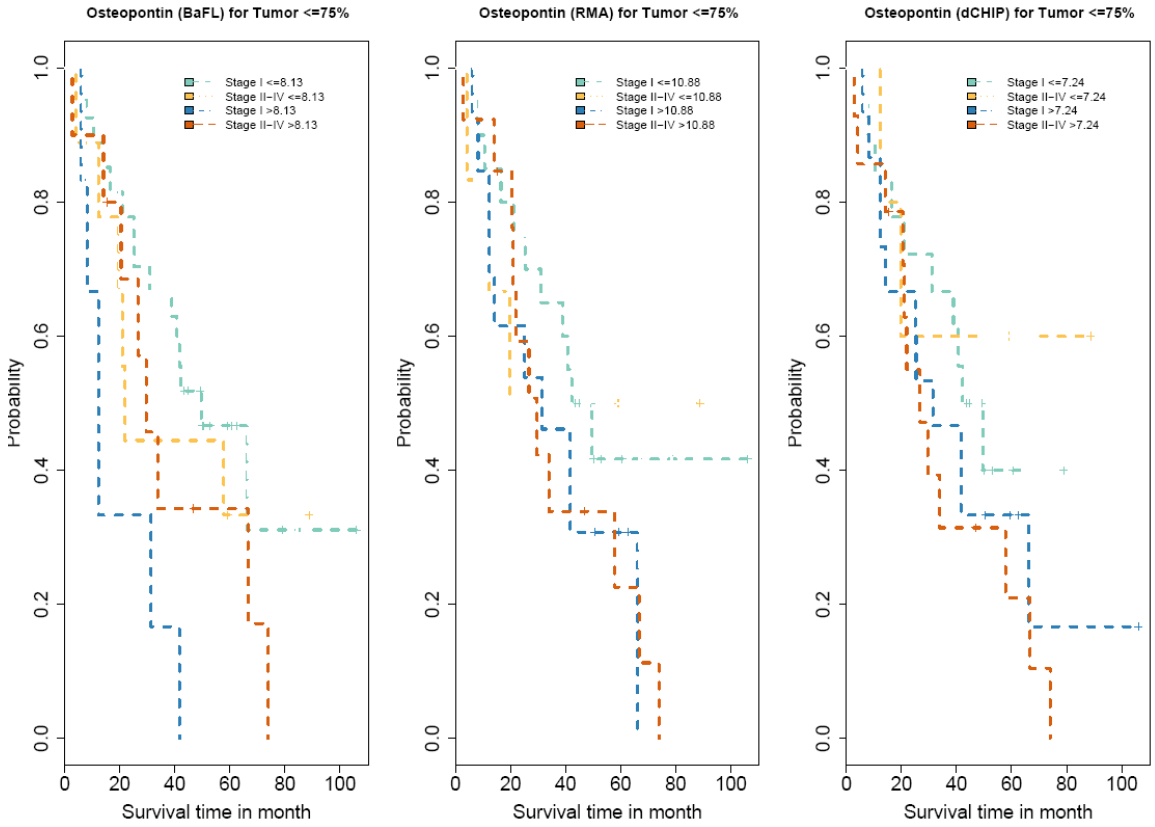


Figure 4.7: Low grade tumor survival curves. ProbeSet 34342_s_at (SPP1) and patient survival rates with respect to less than or equal to 75% tumor pathology and stage progression. The three value-generating methods no longer agree. With the BaFL panel it can be seen that, given low tumor stage histology call but an elevated osteopontin level, there is a significant impact on survival, more so than for late stage cancer patients. The RMA data shows a minor impact, while the dCHIP data show no such effect; the principal difference between RMA and dCHIP is that one sample that was identified as switching groupings in Figure 4.5.

Bonferroni Feature Selection Characteristics

The 30 ProbeSets which were present in the intersection of the Bonferroni corrected t-test results are presented in Table 4.1. In addition to these 30 ProbeSets the osteopontin (SPP1) ProbeSet, 34342_s_at, was included to arrive at the final candidate gene list. This ProbeSet was re-incorporated for the Bhattacharjee normal samples, for the 7 probes retained across all adenocarcinoma samples. This ProbeSet was not eliminated in the BaFL cleansing of the

Stearman data, since 4 probes were retained by the Stearman normal samples. By comparison 13 probes were retained by the Stearman adenocarcinoma samples.

Table 4.1: Final candidate list. The candidate list of 31 genes from the Bhattacharjee and Stearman datasets, as down selected using the Bonferroni correction. This list includes the osteopontin ProbeSet. Note that ProbeSet 34320_at does not align to a specific gene, under the Ensembl genome build used in these studies, although it does align to one transcript region.

ProbeSet	**Hugo Id**
37398_at	PECAM1
40282_s_at	CFD
40841_at	TACC1
36627_at	SPARCL1
37027_at	AHNAK
39066_at	MFAP4
39145_at	MYL9
32562_at	ENG
31856_at	LRRC32
33904_at	CLDN3
33137_at	LTBP4
38995_at	CLDN5
1597_at	GAS6
38704_at	MACF1
895_at	MIF
37658_at	GAS6
36569_at	CLEC3B
39631_at	EMP2
32052_at	KRT121P
39760_at	QKI
770_at	GPX3
34320_at	
33756_at	AOC3
37009_at	CAT
40202_at	KLF9
36155_at	SPOCK2
32847_at	MYLK
38044_at	FAM107A
41338_at	AES
35152_at	RAMP3
34342_s_at*	SPP1*

The candidate gene list values were used as input for cross dataset validation of the classification performance for the three algorithms described in Chapter 3: LDA, kNN, and RF. As described there, the validation was carried out over 100 random permutations of both the training and test datasets [10, 21]. The classification performance of the BaFL interpreted genes is benchmarked against the RMA and dCHIP values input for the same classifiers. Model performance was assessed by the AUC for the classification success [10, 14, 15]; results are presented in Figure 4.8.

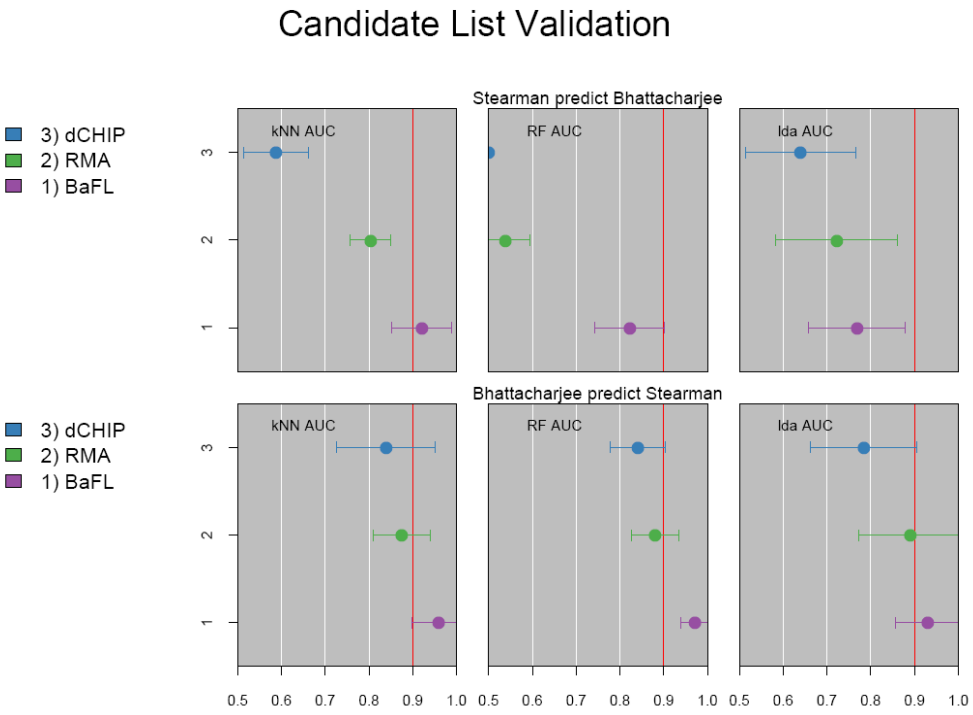


Figure 4.8: Candidate list validation. Classification performance for the candidate list of 31 ProbeSets. These ProbeSets were elucidated using the Bonferroni correction of the t-test results and are the set of genes lying in the intersection between the two datasets along with the osteopontin Probeset. The BaFL based values for these ProbeSets demonstrates the best performance for all three classification algorithms and both reciprocal training and test scenarios.

Discussion

In this chapter we have presented latent structure associated with the genes in our candidate ProbeSet list when BaFL-provided values are used, based upon correlation of the disease samples (Figures 4.1-3). This is the reverse of the normal strategy, in which an analyst uses gene correlations to explore the structure between samples. There is an improvement in the Bhattacharjee data model when the BaFL pipeline is implemented to supply the Probeset values as compared to the RMA and dCHIP methods.

The impact of osteopontin expression levels (Figure 4.4-7) is a well known feature in the progression of lung cancer, as is the difference among lung cancer sub-types [19, 20, 22, 23]. However, a novel result of the analysis is the demonstration that the significance of the expression levels is not stage specific (Figures 4.6-7), but rather pathology specific, for BaFL-supplied values. This change in perspective suggests why our data fails to demonstrate the same survival rates as the Donati study which suggested long term survival of stage I patients with elevated levels of opsteopontin [20]. An attempt to contact this group to obtain supplementary data in order to determine whether their stage I sub-population had a significant proportion of low grade tumor patients has been unsuccessful. Osteopontin is under development as a potential biomarker, and efforts have been made to improve the performance of the biomarker by coupling it with other biomarker results [ref]. Our results suggest that an additional element leading to success would be to incorporate the tumor grade clinical data into the test.

In earlier chapters the multiple-test problem was ignored (leading to the candidate list of 325 ProbeSets described above). The Bonferroni correction was implemented for its low tolerance for

rejecting the null hypothesis [8], and applied to the two datasets as interpreted by the BaFL pipeline. This correction enabled us to elucidate a small subset of undeniably differentially expressed genes as pertinent to lung cancer. The candidate list of genes when based on BaFL pipeline values yields reasonable classification performance across 3 independent algorithms when an appropriately sized dataset is used (Figure 4.8). The size of the data set is important: when the Stearman dataset was used for training purposes all three models struggled, with the BaFL data performing the best in all cases Statnikov, et al. used the Bhattacharjee multiclass data in their pipeline for the cancer diagnosis and biomarker discovery, in which whey reported the maximum prior probability of a dominant diagnostic category of 68.5% [24]. Their analysis used the data in a training approach for 10 fold leave one out cross validation and reported perfect prediction for the training model [24]. Additional analysis utilizing the full Bhattacharjee multiclass dataset yielded binary classification accuracies of 52-56% for random forest and support vector machines with and without gene down selection. While multiclass classificaiton, with and without gene down selection yielded 77-82% for random forest and 89% for support vector machines [25].

The relevance of these 31 genes is supported by the GO connections identified by the pathway and literature search, PaLS, software (http://pals.bioinfo.cnio.es/), which connect 23 of the 30 genes (one ProbeSet aligns to no defined gene). KEGG pathway connections link 6 of these genes, including opsteopontin, through focal adhesion and extra-cellular matrix receptor intreraction [26]. Other extra cellular matrix genes include MFAP4, SPARCL1, ENG, RAMP3, and LRRC32. SPARCL1 and SPOCK or SPARC like have been investigated for their role in lung cancer [23]. These genes along with SPP1, MIF, and PECAM1 have strong immunological associations and thereby may be essential for angiogenesis and tumorogenesis [19, 20, 27]. The

discovery of MACF1, a microtubule-actin crosslinking factor 1, may have associations with TACC through its anchoring to the golgi apparatus and its molecular mobility [28]. There are 3 TACC human genes, which appear to be important for cellular division and organization [28, 29], with orthologues reported in *Mus. musculus*, *Drosophila melangaster*, and *Xenopus laevis*. Proper localization of TACC during cytokinesis appears to be dependent upon phosphorylation by aurora kinases [29, 30] and may posses a critical function in cell cycle control [31]. GAS6, growth arrest 6, is a gamma-carboxyglutamic acid (Gla)-containing protein and contrary to its name is thought to be involved in the stimulation of cell proliferation. Additionally, 3 genes are metabolic enzymes involved in tryptophan or tyrosine metabolism, as presented in Figure 10 [26].

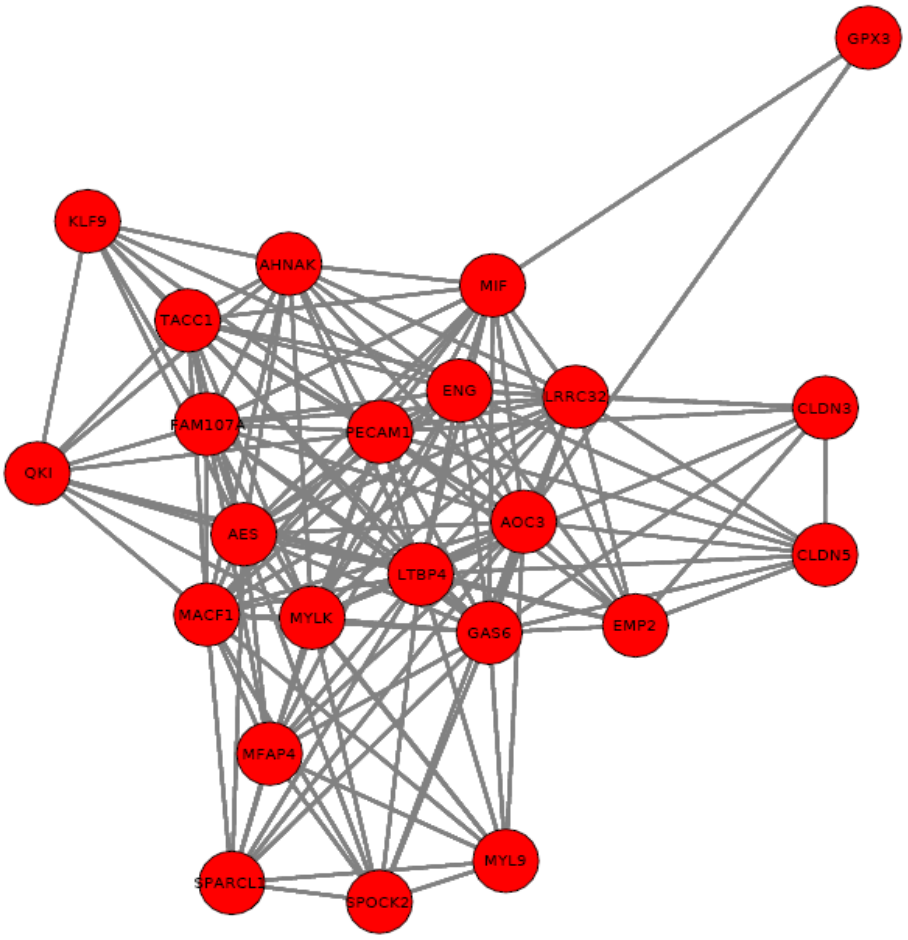


Figure 4.9: GO connectivity of candiadate genes. PaLS software pathway [26] associations of GO terms for the 31 candidate gene that were produced using the BaFl methodology.

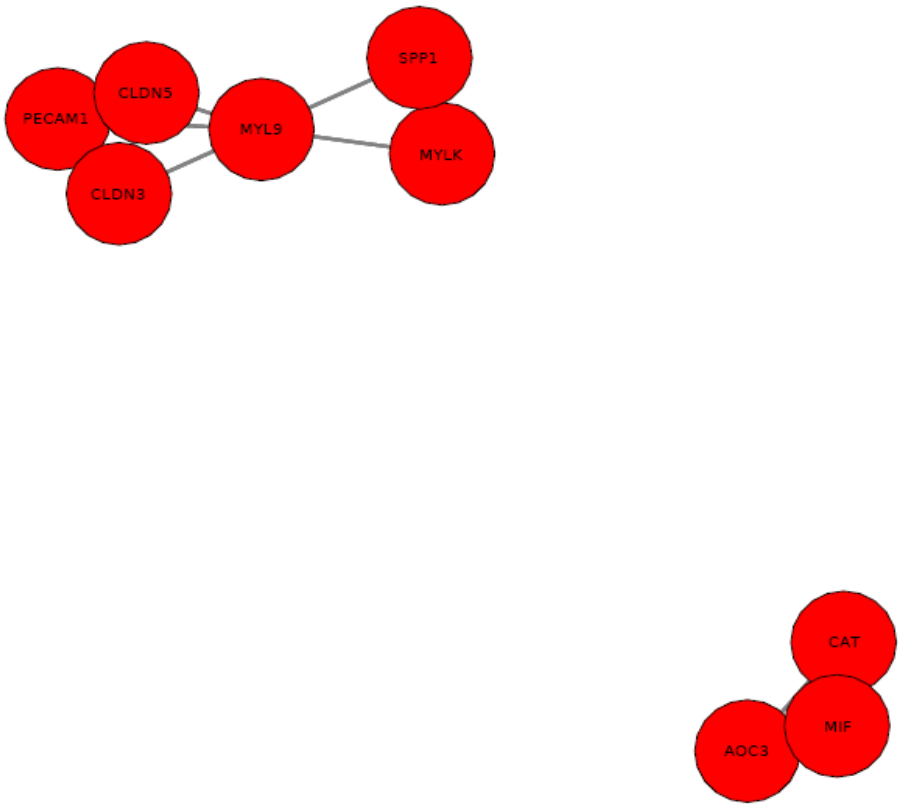


Figure 4.10: KEGG connectivity of candidate genes. PaLS software pathway [26] associations of KEGG terms for the 31 candidate gene that we present here.

Conclusion

We identified 31 genes that yield good cross experiment classification performance between the Bhattacharjee and Stearman datasets. These genes were derived as the intersection of the 30 Bonferroni t-test corrected differentially expressed genes between the two datasets, with the addition of osteopontin. Of these, 30 are a subset of the 325 dataset concordant differentially expressed genes which demonstrate strong separation of regulation patterns in the latent structure

under linear and non-linear dimension reduction techniques. The 1st Laplacian dimensions demonstrate persistence of this structure across the two datasets. In addition to these 30 genes, osteopontin was incorporated as we have demonstrated its significance to tumor progression, which agrees with the literature results for this gene. These 31 genes not only demonstrate significant classification abilities, but they also demonstrate a significant association in the GO terminologies which describe them.

Chapter 5: Data Mining the Multiclass Dataset

Lung cancer is the most prominent form of cancer worldwide, representing 12.3% of all cancer diagnoses [1]. It is a devastating disease, for 90% of those diagnosed with lung cancer will eventually succumb to it, representing 17.8% of cancer deaths worldwide. Smoking tobacco is strongly correlated to the development of lung cancer, with 80-90% of all diagnoses being attributed to smokers, although only 11% of cigarette smokers will develop lung cancer [1]. There exist four histologically distinct lung cancer variants: 3 are of non small cell lung cancer (NSCLC), and the other is small cell lung cancer (SMLC) [1]. NSMLC is the predominant form of lung cancer, encompassing 80% of all lung cancer cases [2]. Adenocarcinoma has surpassed squamous cell carcinoma in prevalence; both are subtypes of NSCLC, the most frequent subtype of lung cancer [1, 2]. Alarmingly, adenocarcinoma is most common in women, non-smokers, and the young [1, 2]. Adenocarcinomas are peripherally located in the lungs and develop from clara cells, alveoli, and mucin producing cells. Squamous carcinomas arise in the central airways and are the direct result of smoking, as there are no squamous epithelial cells in normal lungs. Surgical intervention for patients without mediastinal involvement still results in only a 30-50% chance of disease-free survival, with long-term survival greatly reduced for patients with mediastinal involvement [1].

We have separated the samples from the Bhattacharjee dataset into several subsets, in order to produce a NSCLC multiclass dataset consisting of adenocarcinomas, squamous cell carcinomas, and normal (actually adjacent) biopsies. There exist 155 samples which survived the BaFL

sample cleansing process and these encompass 4,248 ProbeSets. Borrowing from machine learning practices, a gain ratio was calculated from *k*-means clustering analysis of individual ProbeSets to perform down selection to the most significant features [3, 4]. This gain ratio was also calculated at the individual probe level and then averaged across the probes belonging to an accepted set, to give a ProbeSet response. We demonstrate that both approaches improve the model's performance in a supervised classification analysis, implemented using either the kNN classifier based on Euclidean space [3, 5, 6], or Fisher's linear discriminant analysis (LDA) [7, 8]. In this case the emphasis is not on comparing analysis methods, but rather focuses on discovery of intriguing biological phenomenon revealed by using the BaFL pipeline to select the most unambiguous signal. Eighteen significant ProbeSets were selected, possessing a gain criterion greater than 0.8, and this set of genes suggest that an important mechanism underlying tumorogenesis is abnormal cytokinesis. The biological significance of these genes are validated by a literature survey in the discussion section.

Materials and Methods

A subset of samples were selected from the BaFL cleansed Bhattacharjee dataset to construct a multiclass dataset containing NSCLC tumor biopsies and adjacent/normal biopsies. This dataset contained the 125 adenocarcinomas and 13 normals, which comprised the two state disease model previously considered, and an additional 17 squamous samples. The BaFL cleansing pipeline was applied across all samples, with the result that 24,022 probes were found to be common to the three states; these lie in 4,248 ProbeSets (with at least 4 acceptable probes each).

Information gain ratios (equations 2 and 3) were calculated per probe and ProbeSet, based upon their performance as implemented in the k-means clustering algorithm, for three distinct clusters. The average gain ratio for all of the probes in a ProbeSet was determined. However for this analysis, the gain ratio calculated for the average ProbeSet intensity was evaluated to eliminate less informative ProbeSets. It is proposed here that the differences observed between the aggregate ProbeSet and the average probe performance can by utilized in a way similar to the suggestion for using the 'Signal' ProbeSets in Chapter 3, to discern transcript regions relevant to the phenotype. Prior to calculating the gain ratio, normalization transformation was performed on the data (probe and ProbeSet). Let $x_{i,j}$ represent the data as 4,258 ProbeSets by 155 samples and was scaled as

$$x_{i,j} = (x_{i,j} - \bar{x}_i)/\sigma_i . \quad (5.1)$$

Where $\bar{x}$ is the mean signal intensity across the samples and sigma the variance across samples. Hartigan-Wong Clustering was done for 50 random centers (nstarts=50) [9], which appeared to be sufficient to minimize the Euclidean sum of squares. This clustering approach is the default and according to the R documentation it typically demonstrates the best performance [9]. The two best solutions were then chosen, and their gain ratios were calculated, given their distinct cluster centers. These solutions were selected as optimal for the adenocarcinoma clustering, having the smallest Euclidean sum of squares and sub-optimal for either the squamous or normal clustering. The decision to use both best solutions was an effort to compensate for the 'no free lunch theory' [10], in that if the clustering was appropriate for both adenocarcinoma and normal samples, the clustering underperforms for the squamous samples. Fifty clustering attempts were typically sufficient to find both 'best' centers, and the gain ratios for the two centers were

averaged to give a final approximation of the gain ratio; where the randomness of initial selection prevented the discovery of both solutions the single solution gain ratio was used. The gain ratio was calculated as such:

$$gainratio(X) = gain(X) \Big/ splitin(X), \text{ where} \tag{5.2}$$

$$splitin(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} * \log_2\left(\frac{|T_i|}{|T|}\right) \tag{5.3}$$

Here T_i represents the number of classification calls in each cluster [4]. As individual ProbeSet's inherent clustering ability increases, the information gain ratio approach 1.0 [3, 4, 10]. Down-selection to the most informative ProbeSet features was set by a gain ratio threshold criterion. For this dataset gain ratio limits of 0.7, 0.8, 0.9 yield ProbeSet lists of length 43, 18 and 8, respectively. Average probe gains greater than or equal to 0.5, 0.6, and 0.7 yield ProbeSets lists of length 95, 29 and 9.

Performance of the gain criterion down selected ProbeSets was monitored with supervised kNN and LDA classification performances. Training and testing was done under 100 X 2 validations with random splitting of samples in training and test sets, with sampling done with replacement [3, 11, 12]. Sampling with replacement in such a scenario allows for the stochastic elimination and pseudo-replication of samples, although replication may have not been consistent to the same training/test set. Majority voting was implemented for the kNN algorithm, with the analysis of 3 nearest neighbors. The area under the receiver operating curve (AUC) was weighted for each of

the pairwise 2-class classification models, and the accumulated AUC represents the multiclass performance [3, 10, 13, 14]. The weighting scheme was as follows:

$$AUC = \sum auc * (1- (C_i/T))/n-1, \quad (4)$$

where C is the class numbers, T is the sample numbers, and n is the number of classes. The weighting scheme was chosen over the average AUC, to compensate for the large class bias. Classifying the adenocarcinoma samples with little classification ability for the smaller classes yielded an inflated model AUC.

Results

Eighteen genes are implicated for differentiating NSCLC when the gain criterion threshold is set at 0.8. These ProbeSets are provided in Table 4.1, along with their HUGO identification, chromosome location, the calculated gain ration and the GO processes. Investigation of the gene's GO processes includes cell proliferation, mitotic cell cycle control, cell motility and adhesion, inflammatory response, signal transduction, and cell programmed death.

Table 5.1: NSCLC candidate genes. The list of 18 genes from the Bhattacharjee NSCLC dataset, as down selected using the gain criteria. *The probes from probeset_id, 576_at, measure the chromosome 7 transcript region 150,342,056-150,342,499, which has overlapping opposed sense genes for ATG9B (ATG9 autophagy related 9 homolog B) and NOS3 (nitric oxide synthase 3 -endothelial cell)

Probeset_id	Gene_id	Chromosome	Gain	GO process
40841_at	TACC1	8	1.018	Cell cycle, cell division
34294_at	KIFC3	16	0.992	Golgi organization and biogenesis, microtubule-based movement, visual perception
39631_at	EMP2	16	0.988	cell proliferation
576_at*	ATG9B	7	0.957	autophagic vacuole formation, autophagy
576_at*	NOS3	7	0.957	angiogenesis, cell motility, learning, lipopolysaccharide-mediated signaling pathway, lung development, negative regulation of calcium ion transport, negative regulation of hydrolase activity, negative regulation of potassium ion transport, negative regulation of smooth muscle cell proliferation, nitric oxide biosynthetic process, ovulation from ovarian follicle, oxidation reduction, regulation of sodium ion transport, signal transduction
37004_at	SFTPB	2	0.954	lipid metabolic process, organ morphogenesis, regulation of liquid surface tension, respiratory gaseous exchange, sphingolipid metabolic process
39016_r_at	KRT6A	12	0.935	cell differentiation, ectoderm development, positive regulation of cell proliferation
39066_at	MFAP4	17	0.920	cell adhesion, signal transduction
1718_at	ARPC2	2	0.894	cell motility, regulation of actin filament polymerization
33756_at	AOC3	17	0.883	amine metabolic process, cell adhesion, inflammatory response, oxidation reduction
32052_at	KRT121P	11	0.861	keratin 121 pseudogene
33323_r_at	SFN	1	0.858	DNA damage response, signal transduction resulting in induction of apoptosis, apoptotic program, cell proliferation, keratinocyte differentiation, negative regulation of caspase activity, negative regulation of protein kinase activity, regulation of cyclin-dependent protein kinase activity, release of cytochrome c from mitochondria, signal transduction, skin development
37009_at	CAT	11	0.855	UV protection, hydrogen peroxide catabolic process, negative regulation of apoptosis, oxidation reduction, protein tetramerization, response to reactive oxygen species
36495_at	FBP1	9	0.842	carbohydrate metabolic process, fructose metabolic process, gluconeogenesis
34301_r_at	KRT17	17	0.841	biological_process, epidermis development
32680_at	TNIK	3	0.832	JNK cascade, protein amino acid phosphorylation, protein kinase cascade, response to stress
41639_at	NCAPH	2	0.832	cell division, mitosis, mitotic cell cycle, mitotic chromosome condensation
654_at	MXI1	10	0.825	cytoplasmic sequestering of transcription factor, negative regulation of cell proliferation, regulation of transcription, regulation of transcription, DNA-dependent

Baseline comparison of the weighted AUC (equation 4) to the 4,248 ProbeSets are made in Figure 5.1. The feature subsets of lengths 43, 18, and 8 were selected for possessing gain criteria greater than: 0.7, 0.8, and 0.9, respectively. The classification performance for these candidate ProbeSet was assessed for the kNN and LDA classification algorithms [5-8]. The ProbeSets selected by their gain criterion demonstrate classification performance improvement against the whole model's classification ability, and more significantly so for the kNN algorithm. While the most ProbeSets identified by the most stringent threshold (greater than 0.9 information gain) demonstrate a distinct loss of classification ability. In general, the linear discriminant analysis demonstrates a marked improvement over the kNN classification algorithm, and not as significant an improvement by down selection.

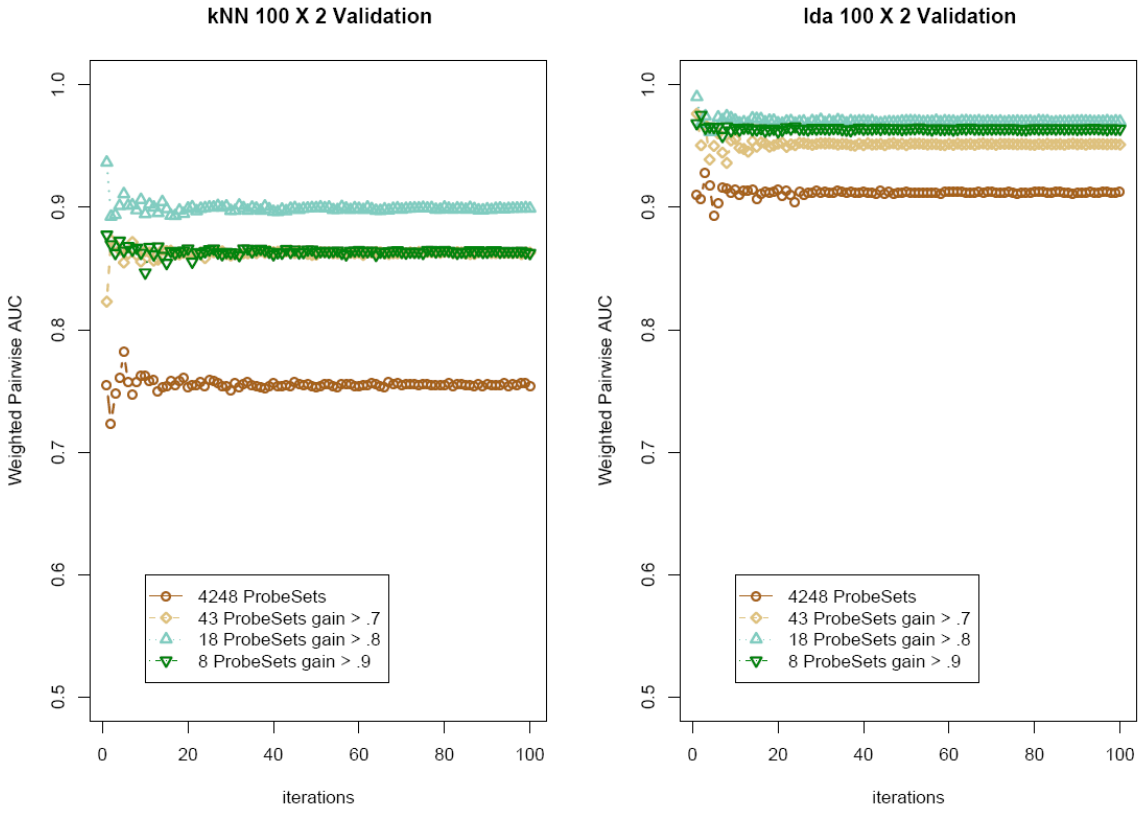


Figure 5.1: NSCLC ProbeSet gain selection model performances. Down selection of ProbeSets by the ProbeSet gain criterion, as presented by the weighted AUC metric. Gain criterion was calculated for the k-means clustering performance of the average ProbeSet intensity. Left graphic is kNN implemented with k=3 and majority voting and right graphic is LDA.

Down selection by the gain criterion calculated as the average probe clustering performance across the ProbeSet yields similar classification performance for the LDA algorithm, and marked improvement in the kNN classification algorithm. Results of the down selection through the average probe gain criterion are presented in Figure 5.2. However, this is questionable approach since the ProbeSet aggregate may possess minimal gain relevance. It is our belief that the ProbeSets which are identified by both the ProbeSet and average probe gain criteria act as strong classifiers, while the less informative ProbeSets which were identified by their average probe information gain act as weak to noise classifiers.

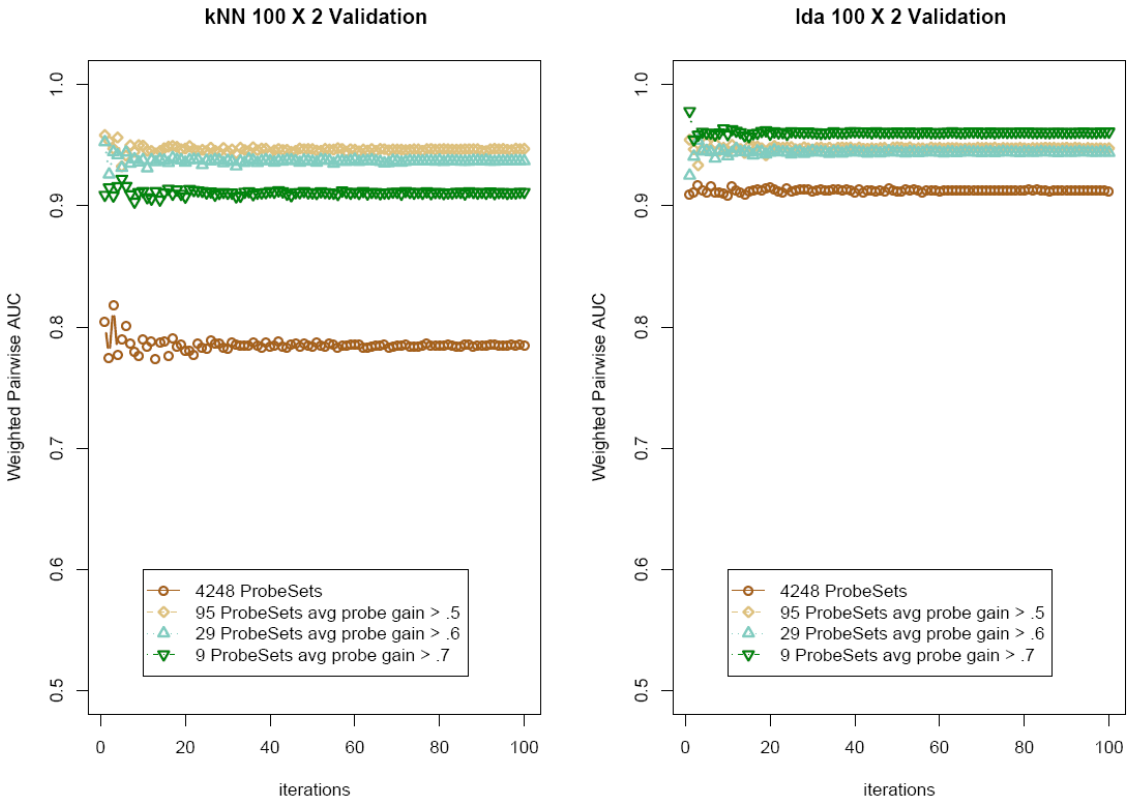


Figure 5.2: NSCLC average probe gain selection model performances. Down selection of ProbeSets by the average probe gain criterion, as presented by the weighted AUC metric. Gain criterion was calculated for the k-means clustering performance of the average ProbeSet intensity. Left graphic is kNN implemented with k=3 and majority voting and right graphic is LDA.

A more suitable approach is to identify the ProbeSets meeting an average probe gain criteria, which is still less then the aggregated ProbeSet's information gain. We identified 13 such ProbeSets, whose average Probe information gain was greater or equal to 0.6 and the aggregated ProbeSet gain exceeded this threshold. The classification performance of these ProbeSets is presented in Figure 5.3, again in comparison to the whole data model.

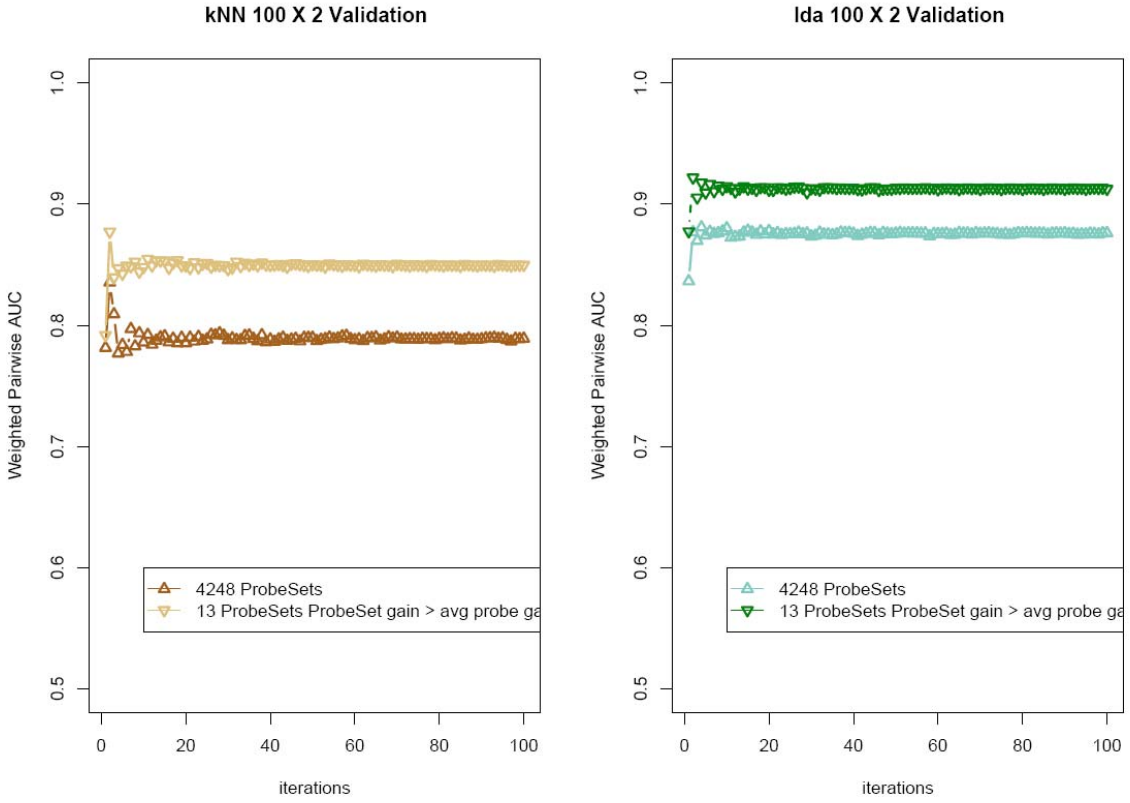


Figure 5.3: NSCLC refined average probe gain selection model performances. Classification performance of kNN (top) and Fisher's linear discriminant analysis (bottom), as demonstrate by the weighted AUC per individual disease class. Left column of graphs presents the performance of the full 4,248 ProbeSets and right column of graphs present s the performance of the 13 ProbeSets with and average gain >= 0.6 with the aggregate gain greater than average gain.

Discussion

Significant improvement in sample classification is demonstrated, with all three down selected ProbeSets lists, for both classifiers. The linear discriminant analysis of the full model demonstrates exceptionally good classification properties, much moreso than that of the kNN algorithm. However the performance of the LDA algorithm is still improved through the information gain down selection criterion. While the results from using most stringent gain criterion, 0.9, indicate some over-training of the model, since both classifiers demonstrate

decrease performance in the weighted AUC metric. This list of candidate genes demonstrate a decrease true positive classification rate for the squamous samples, with coordinating increases in false positive rates for the adenocarcinoma and normal samples (data not shown).

Although ProbeSets selected by their average probe gain criteria demonstrate enhanced performance for the kNN algorithm and similar performance for the LDA algorithm, it likely not a suitable approach. We observed little overlap (6) between the 29 ProbeSets selected by their average probe gain (>=0.6) and the 18 ProbeSets selected by their ProbeSe5t gain criteria (>=0.8). This translates to only $1/5^{th}$ of th ProbeSets having an enhanced information gain for the aggregated ProbeSet, and in fact $2/3^{rd}$ of them demonstrate diminished information gain for the aggregated ProbeSet. Therefore, we believe that the enhanced classification performance for the average probe information gain selected lists is likely the result of the 'good' classifiers being more significant in the feature subspace.

However the average probe gain suggests probes to investigate for transcript measurement anomalies. For example, the same list of 9 ProbeSets included 6 ProbeSets containing distinct probes demonstrating poor clustering performance, and these have been show to skew the gain distributions and presumably the aggregate's clustering performance. For example, the KRT17 ProbeSet has 2 probes (index 6 and 10) with low probe performances, 0.033 and 0.06 gain ratios respectively. Selection by the average probe gain per ProbeSet appears to have a high false positive rate; furthermore these ProbeSets appear to have limited connection to the ProbeSets selected by the ProbeSet gain criterion. As a counterexample, ProbeSet 33108_i_at, SOX2, includes five probes all of whose gain values range from 0.73 to 0.99, yet the ProbeSet aggregate gain ratio value is a dismal 0.18. This ProbeSet is up-regulated in the squamous samples, by the

statistical criterion, but it appears to be unable to classify the adenocarcinomas because the Euclidean Sum of Squares is two fold larger than for the (smaller N) squamous and normal clusters. The ProbeSets KRT17, EMP2 and MFAP4 which were selected by *both* criteria, demonstrate that one method by which to avoid such false positives: that is, select an average probe gain threshold criteria of ProbeSets whose aggregate gain improves on the average probe gain. Thirteen of 29 ProbeSets demonstrating an average probe gain had an aggregate gain ratio that improved on the average probe gain ratio. In contrast to the average probe gain list of 29 genes, half of this list (6 of 13) demonstrates enhanced ProbeSet information gain; having a ProbeSet gain ratio greater than 0.8. The 7 additional genes for demonstrate a close relationship to the underlying biology.

Such effective down selection, at the ProbeSet level, may be a reasonable endpoint if the goal is to produce an effective and simple diagnostic test. However, this is a limited approach if the goal is to understand the biological mechanisms of disease. Based upon the classification performances, presented in Figure 5.1, 18 Genes were identified were selected as a candidate list (Table 5.1) to explore the biological function with respect to NSCLC.

Literature Validation of Biological Process

We observed significant down regulation of TACC1 in both forms of NSCLC, which is supported by the literature [15-17]. There are 3 TACC human genes, which appear to be important for cellular division and organization [18], with orthologues reported in *Mus. musculus*, *Drosophila melangaster*, and *Xenopus laevis [18]*. Proper localization of TACC during cytokinesis appears to be dependent upon phosphorylation by aurora kinases [15, 19] and may posses a critical function in cell cycle control [17]. Appropriate cell division is a prerequisite in the propagation

of cells, and abnormal cytokinesis can lead to spontaneous abortions, congenital malformations, polyploidy and cancer predisposition [15, 19].

Similarly, EMP2 was significantly down regulated in the NSCLC samples. EMP2 has been demonstrated to down-regulate caveolin-1 production [20], with up-regulated caveolin-1, CAV1, inducing filopodia and facilitating metastasis in lung adenocarcinoma [21]. Caveolin-1 is essential for the formation of caveoli formation, the plasma membrane invaginations which alter the morphology of the plasma membrane [20]. The invaginations are essential for cholesterol transport, localization of signal transduction, coordination with microtubules for membrane trafficking, and regulation of NOS3, while oncogenic implications have been the subject of conflicting reports for different cancer types [20, 22-24]. The caveolae structures constantly recycle between the cell-to-cell contact points along the cell membrane, endosomes, and Golgi network, while during mitosis they localize along the contractile ring [23, 25].

Additionally, the DNA-damaging ROS H_2O_2 is a mitotic signal messenger [26], which catalase rapidly degrades [27]. Catalase (CAT) was demonstrated to accelerate the degradation of p53 proteins, thereby preventing ceramide induced apoptosis [27], while IGF-1 was demonstrated to restore catalase activation through inhibition of PI-3-Akt inhibition at CAV1 localized microdomains [28]. We observed significant upregulation of CAT in the normal samples.

Similarly, stratifin is induced by DNA-damage via p53 activation and regulates G2 cell cycle arrest through positive feedback upon p53 [29, 30]. Conversely, it has been reported the stratifin is activated by IGF1 in a manner independent of p53 activation [31]. Methylated sequesence of stratifin has been reported in several cancers [30] and improper regulation may explain cancers

lacking p53 mutations [30]. While, we observed significant up-regulation of stratifin in the Bhattacharjee squamous samples, the adenocarcinoma samples demonstrated minimal presence of p53 mRNA. It is our belief that these gene levels suggest that a flawed cytokinesis process is occurring in these samples. Further support of this are the presence of ARPC2 and TNIK, both of which regulate the actin cytoskeleton [32-34].

These results may also indicate that the squamous samples are more proficient at attaining appropriate G2 sequesence, although in the prolonged G2 state they may be incurring an accumulation of ROS, causing DNA-damage. Additional support for the occurrence of DNA-damage is the association of condensing-1 with PARP (poly ADP-ribose polymerase), supporting the role of condensing-1 in DNA repair [35]. Similarly, TTF-1 has been shown to interact with PARP2 to regulate the expression of surfactant protein B [36].

Surfactant protein B (SPB) and keratin 6A both were significantly up-regulated in the squamous samples and slightly up-regulated in the adenocarcinoma samples. Surfactant protein B is a 79 amino acid hydrophobic peptide which is expressed in alveolar type epithelial cells and clara cells of the lung [37-41]. This peptide is essential in maintaining normal lung functions and surface membrane structure, by reducing the surface tension [37]. The up-regulation of this peptide may be a compensation for the toxic exposure of lung tissue to cigarette smoke: the link to TTF1 is tantalizing. TTF1's DNA binding activity of SPB is well documented [36, 38, 39]. Interference of TTF1 binding has been reported to be caused by ceramide [39], PARP-2 [36], and proteosome dysfunction [38]. Coupled histological staining for TTF1, KRT6A, and p63 has recently been evaluated for NSCLC classification [42]; additionally, p63 and KRT6A have been associated with more aggressive tumors [43].

Table 5.2: Genes identified through refined average probe gain. The list of 8 genes from the Bhattacharjee NSCLC dataset, as down selected using the refined average probe gain criteria.

probeset_id	Gene_id	chromosome	gain	avg probe gain
1794_at	CCND3	6	0.687	0.605
31775_at	SFTPD	10	0.743	0.652
32052_at	KRT121P	11	0.861	0.697
33109_f_at	SOX2	3	0.800	0.628
36617_at	ID1	20	0.701	0.694
36629_at	TSC22D3	X	0.610	0.607
37926_at	KLF5	13	0.683	0.631
40786_at	PPP2R5C	14	0.685	0.678

The 8 ProbeSets identified by the refined selection of ProbeSets with sufficient average probe gain to pass the double selection criterion complement the underlying biological story developed above. ID1 is a member of the basic helix-loop-helix transcription factors which regulate a range of cellular functions, including cell cycle progression/proliferation [44, 45]. ID1 regulation has been linked to several pathways, including matrix metalloproteinases [45] and bone morphogenetic proteins [46]. It has been demonstrated that CAV1 binds ID1 through the HLH domain and this interaction was essential for the activation of AkT activation, promoting cancer cell invasion in prostate cancer [45]. Additionally, ID1 has been demonstrated to be essential for the expression of cyclin D1, a facilitator of cell phase transitions [44]. While the cyclin D3 ProbeSet (1794_at) demonstrated a gradient transition of expression levels between normals, squamous, and adenocarcinoma samples, the highest mean expression was in squamous cell carcinomas. Cyclin D3 plays a pivotol role in the phase transition between G1 to S stages [47]. The adenocarcinoma samples similarly expressed significant down regulation of the ID1 mRNA transcript, suggesting that these samples are residing in the growth stages.

Kruppel-like factor 5 (KLF5) has also been implicated in carcinogenesis pathways. In particular, it activates in response to DNA damage to induce apoptosis or DNA repair, through the p53 dependent pathways [48]. Increased KLF5 expression accelerates G2/M phase transitions by activating cyclin B1 [49]. Similarly, PPP2R5C also participates in p53 activation after the induction of DNA damage, acting as a mitotic checkpoint to suppress cancer growth [50, 51]. DNA damage activates ATM (ataxia telangiectasia mutated), which phosphorylates p53 at Ser15. The phosphorylation promotes the p53 and PPP2R5C interaction, thereby causing the dephosphorylation of p53-Thr55 and cell cycle arrest [50, 51]. Both KLF5 and PPP2R5C have elevated levels of mRNA transcript in the squamous cells, which supports the hypothesis that the squamous carcinomas have undergone DNA damage. Finally the TSC22D3, or GILZ, ProbeSet demonstrates significant up-regulation in the normal samples. This gene regulates a number of cell functions, including cell proliferation, and has been identified has a tumor suppressor gene.

Conclusion

Cancer is a complex disease that can affect nearly every pattern of expression of a cells genes, depending on the stage. This obscures important mechanisms that discriminate the types of disease and perhaps possible points of control. We have shown that by application of the BaFL cleansing routine, followed by straightforward machine learning methods for down-selection criteria and clustering parameters, even very simple classification methods reveal a distinct picture of tumorogenesis for the Bhattacharjee NSCLC samples. While the complexity of the disease cannot be overstated, this picture is much simpler and amenable to straightforward assays for confirmation. For example, a significant proportion of the identified genes have been documented to have alternative transcripting events, oncogenic genomic amplification,

methylation regulatory control, and multiple cellular functions, for which well documented methods exist. We believe that the BaFL cleansing process has provided the basis for an intriguing elucidation of distinct differences in disease mechanism for two histologically distinct NSCLC lung tissue tumor types.

Appendix A

This is the main driver function in the CleansingPrep_6_23.py file which prepares the ProbeFATE system for the BaFL filter cleansing

tables parameter is a list of files to import or create
example files are in the Supplementary Materials:

1) MissingQuantitation.csv quantitation types which I add to the system
2) Known.csv calculates or uses previously calculated deltaG (OligoArrayAux) values for probes with known sequences, flag meth[0] differentiates
3) SampleMask.csv a file which describes the data (abbrev. sample sd, orig sample id, the associated .cel file, disease class, and the file source for this information
4) SNP_by_pos.csv a file created w/ Sunita.py, which parses the information from AffyMAPSDetector output
5) probe_mapping.csv file is created through ENSEMBL mappings or probes, flag[1] indicates whether the latest ENSEMBL build is used otherwise the build needs to be provided

```
def prepareCleansing(db, usr, pswd, fpath, logfile, tables, meth):
      # tables = [added_qts, sample_info, snp_info, xhybrid_info]
      #  add new qts
      i=0
      CreateNew_QTs(db, usr, pswd, tables[0], logfile)
      i+=1
      # map known chip sequence info
      if lower(meth[0])=='copy':
            Copy_deltaG(db, usr, pswd, tables[1], logfile)
            i+=1
      elif lower(meth[0][:4])=='calc':
            Calc_deltaG(db, usr, pswd, fpath, logfile)
      else:
            print 'Method:', meth[0], 'does not exist.'
            return -1

      # set sample mask
      Sample_mask(db, usr, pswd, logfile, tables[i])
      i+=1
      # map SNP info
      SNP_mask(db, usr, pswd, tables[i], logfile)
      i+=1
      # map xhybridization info
      if lower(meth[1])=='new':
            new_Xhybrid(db, usr, pswd, tables[i], logfile)
      else:
            version_Xhybrid(db, usr, pswd, tables[i], logfile, meth[1])
      Mod_wk_registry(db, usr, pswd)
      print '\nXHYBRID INFO MAPPED\n'
      print '\n\tALMOST READY TO PROBE CLEANSE\n'
```

Appendix B

BaFL probe cleansing from the DataCleansing_6_26.py file

```
def RunCleanse(usr, pswd, db, logfile, lwr=200, uppr=20000, rgr=4):
      Driver(usr, pswd, db, logfile, lwr=200, uppr=20000)
      DriverSNP(usr, pswd, db, logfile)
      DriverXH(usr, pswd, db, logfile)
      DriverBioPhy(usr, pswd, db, logfile)
      DriverFNL(usr, pswd, db, logfile, rgr)

def exec_mapping(usr, pswd, db, tables, msk, msk_tab, fp, lwr, uppr):
      cur, conn= make_connect(usr, pswd, db)
      exp=get_exp(usr, pswd, db)
      notes='known probe sequence info within linear range'
        map=get_masking(tables, msk, msk_tab)
        chip_info = " as select
known_seq_biophysical.probeset_id,known_seq_biophysical.probe_index,kno
wn_seq_biophysical.pm_mm_other, known_seq_biophysical.probeseq,"
      k=0
      #print 'Map length ', len(map),     map[k]
      pid_tbles=[]
      for i in tables:
            tmp=msk[map[k]]+'_pid'
            trl= "CREATE table
"+tmp+chip_info+i+".x,"+i+".y,"+i+".signalrawintensity from
"+i+",known_seq_biophysical where known_seq_biophysical.x = "+i+".x and
known_seq_biophysical.y = "+i+".y and (signalrawintensity >= "+
str(lwr) +" and signalrawintensity <= "+ str(uppr) +" and pm_mm_other
>=0)"
            print trl
            cur.execute(trl)
            conn.commit()
            #print trl, '\n\n'
            fp.write('\tTable '), fp.write(tmp), fp.write('
created.\n')
            pid_tbles.append(tmp)
            cur, conn=UpdateWorkReg(cur, conn, tmp, exp,
'linear_range', usr, i, 'known_seq_biophysical', notes)
            k+=1
      onto='select * from '+tmp
      cur.execute(onto)
      desc=cur.description
      conn.close()
      fp=UpdateOntology2(usr, pswd, db, desc, exp, 'linear_range',
'_pid', notes, str(lwr)+', '+str(uppr), fp)
      fp.close()
```

```
def Driver(usr, pswd, db, logfile, lwr=200, uppr=20000):
      # first step when combines chip and cel info
      cur, conn= make_connect(usr, pswd, db)
      tables=get_tables(cur)
      msk, msk_tab, Lmsk, state=get_mask2(cur)
      #print len(msk), len(tables), len(msk_tab)
      t = datetime.datetime.now()
      EpochSeconds=time.mktime(t.timetuple())
      now = datetime.datetime.fromtimestamp(EpochSeconds)
      fp=open(logfile, 'a')
      fp.write('\tTABLE CREATIONS LOGFILE\n\n')
      fp.write(now.ctime())
      fp.write('\nChip to .cel file mappings, -Missing Seq, +Linear
Range:\n')
      conn.close()
      exec_mapping(usr, pswd, db, tables, msk, msk_tab, fp, lwr, uppr)

def snp_filter(usr, pswd, db, msk, msk_tab, fp):
      cur, conn= make_connect(usr, pswd, db)
      exp=get_exp(usr, pswd, db)
      notes='linear range samples w/o SNPs'
      k=0
      #print 'Map length ', len(map),     map[k]
      pid_tbles=[]
      #fp=open(logfile,'a')
      for i in msk:
            trl= 'CREATE table ' +msk[k]+ '_snp as select '
            trl=trl+ msk[k] +'_pid.x, '+ msk[k] +'_pid.y from '+
msk[k]+ '_pid except select snp_list.x, snp_list.y from snp_list'
            cur.execute(trl)
            conn.commit()
            #print trl, '\n\n'
            fp.write('\tTable '), fp.write(msk[k]+ '_snp '), fp.write('
created.\n')
            pid_tbles.append(msk[k]+ '_snp')
            cur, conn=UpdateWorkReg(cur, conn, msk[k]+'_snp', exp,
'snp_filter', usr, msk[k]+'_pid', 'snp_list', notes)
            k+=1
      onto='select * from '+msk[k-1]+'_snp'
      cur.execute(onto)
      desc=cur.description
      conn.close()
      fp=UpdateOntology2(usr, pswd, db, desc, exp, 'snp_filter',
'_snp', notes, '', fp)
      fp.close()

def DriverSNP(usr, pswd, db, logfile):
      # first step when combines chip and cel info
      cur, conn= make_connect(usr, pswd, db)
      msk, msk_tab, Lmsk, state=get_mask2(cur)
      #print len(msk), len(tables), len(msk_tab)
      t = datetime.datetime.now()
      EpochSeconds=time.mktime(t.timetuple())
      now = datetime.datetime.fromtimestamp(EpochSeconds)
```

```
      fp=open(logfile, 'a')
      fp.write('\n')
      fp.write('SNP filter:\n')
      conn.close()
      snp_filter(usr, pswd, db, msk, msk_tab, fp)

def XH_filter(usr, pswd, db, msk, msk_tab, fp):
      cur, conn= make_connect(usr, pswd, db)
      exp=get_exp(usr, pswd, db)
      notes='probes w/o xhybrid'
      k=0
      #print 'Map length ', len(map),     map[k]
      pid_tbles=[]
      #fp=open(logfile,'a')
      for i in msk:
            trl= 'CREATE table tmp as select count(*), probeset_id,
xhybrid_list.x, xhybrid_list.y from xhybrid_list inner join '
            trl=trl+ msk[k] +'_snp on ( '+ msk[k] +'_snp.x =
xhybrid_list.x and ' +msk[k]+ '_snp.y = xhybrid_list.y) '
            trl=trl+ 'group by probeset_id, xhybrid_list.x,
xhybrid_list.y having count(*) = 1 order by probeset_id,
xhybrid_list.x, xhybrid_list.y'
            cur.execute(trl)
            conn.commit()
            trl= 'create table ' +msk[k]+'_XH as select xhybrid_list.*
from xhybrid_list inner join tmp on (tmp.x = xhybrid_list.x and tmp.y =
xhybrid_list.y) order by probeset_id,  xhybrid_list.x, xhybrid_list.y'
            cur.execute(trl)
            conn.commit()
            fp.write('\tTable '), fp.write(msk[k]+ '_XH '), fp.write('
created.\n')
            pid_tbles.append(msk[k]+ '_XH')
            trl='drop table tmp'
            cur.execute(trl)
            conn.commit()
            cur, conn=UpdateWorkReg(cur, conn, msk[k]+'_XH', exp,
'xhybrid_filter', usr, msk[k]+'_snp', 'xhybrid_list', notes)
            k+=1
      onto='select * from '+msk[k-1]+'_XH'
      cur.execute(onto)
      desc=cur.description
      conn.close()
      fp=UpdateOntology2(usr, pswd, db, desc, exp, 'xhybrid_filter',
'_XH', notes, '', fp)
      fp.close()
```

```
def DriverXH(usr, pswd, db, logfile):
      # first step when combines chip and cel info
      cur, conn= make_connect(usr, pswd, db)
      msk, msk_tab, Lmsk, state=get_mask2(cur)
      #print len(msk), len(tables), len(msk_tab)
      t = datetime.datetime.now()
      EpochSeconds=time.mktime(t.timetuple())
      now = datetime.datetime.fromtimestamp(EpochSeconds)
      fp=open(logfile, 'a')
      fp.write('\n')
      fp.write('Xhybrid filter:\n')
      conn.close()
      XH_filter(usr, pswd, db, msk, msk_tab, fp)

def BioPhy_filter(usr, pswd, db, msk, msk_tab, fp):
      cur, conn= make_connect(usr, pswd, db)
      exp=get_exp(usr, pswd, db)
      notes='probes w/o structural issues'
      k=0
      #print 'Map length ', len(map),     map[k]
      pid_tbles=[]
      #fp=open(logfile,'a')
      for i in msk:
            trl= 'create table tmp as select known_seq_biophysical.*
from known_seq_biophysical inner join '
            trl=trl+ msk[k]+ '_XH on ( '+msk[k]+'_XH.x =
known_seq_biophysical.x and '+msk[k]+'_XH.y =  known_seq_biophysical.y)
where dgss>-3.6 and dgss<10000000000  order by
known_seq_biophysical.probeset_id,  known_seq_biophysical.probe_index,
'+msk[k]+'_XH.x, '+msk[k]+'_XH.y'
            cur.execute(trl)
            conn.commit()
            trl= 'create table ' +msk[k]+'_BioP as select
known_seq_biophysical.probeset_id, known_seq_biophysical.probe_index,
known_seq_biophysical.x, known_seq_biophysical.y,
known_seq_biophysical.pm_mm_other, known_seq_biophysical.dgss,
known_seq_biophysical.probeseq, '
            trl=trl+msk[k]+'_XH.chromosome, '
            trl=trl+msk[k]+'_XH.commence, '
            trl=trl+msk[k]+'_XH.finish, '
            trl=trl+msk[k]+'_XH.strand, '
            trl=trl+msk[k]+'_pid.signalrawintensity from
known_seq_biophysical inner join tmp on (tmp.x =
known_seq_biophysical.x and tmp.y = known_seq_biophysical.y) inner join
'
            trl=trl+msk[k]+'_XH on (tmp.x = '+msk[k]+'_XH.x and tmp.y =
'
            trl=trl+msk[k]+'_XH.y) inner join '+msk[k]+'_pid on (tmp.x
= '
            trl=trl+msk[k]+'_pid.x and tmp.y = '+msk[k]+'_pid.y) order
by known_seq_biophysical.probeset_id,
known_seq_biophysical.probe_index'
            cur.execute(trl)
            conn.commit()
```

```
            fp.write('\tTable '), fp.write(msk[k]+ '_BioP '),
fp.write(' created.\n')
            pid_tbles.append(msk[k]+ '_BioP')
            trl='drop table tmp'
            cur.execute(trl)
            conn.commit()
            cur, conn=UpdateWorkReg(cur, conn, msk[k]+'_BioP', exp,
'biophysical_filter', usr, msk[k]+'_XH', 'known_seq_biophysical',
notes)
            k+=1
      onto='select * from '+msk[k-1]+'_BioP'
      cur.execute(onto)
      desc=cur.description
      conn.close()
      fp=UpdateOntology2(usr, pswd, db, desc, exp,
'biophysical_filter', '_BioP', notes, '', fp)
      fp.close()

def DriverBioPhy(usr, pswd, db, logfile):
      # first step when combines chip and cel info
      cur, conn= make_connect(usr, pswd, db)
      msk, msk_tab, Lmsk, state=get_mask2(cur)
      #print len(msk), len(tables), len(msk_tab)
      t = datetime.datetime.now()
      EpochSeconds=time.mktime(t.timetuple())
      now = datetime.datetime.fromtimestamp(EpochSeconds)
      fp=open(logfile, 'a')
      fp.write('\n')
      fp.write('Biophysical filters:\n')
      conn.close()
      BioPhy_filter(usr, pswd, db, msk, msk_tab, fp)

def Fnl_filter(usr, pswd, db, msk, msk_tab, fp, rgr):
      cur, conn= make_connect(usr, pswd, db)
      exp=get_exp(usr, pswd, db)
      notes='probesets w/ statistical rigor'
      k=0
      #print 'Map length ', len(map),     map[k]
      pid_tbles=[]
      #fp=open(logfile,'a')
      for i in msk:
            trl= 'create table tmp as select count(*), probeset_id from
'
            trl=trl+ msk[k]+ '_BioP group by probeset_id having
count(*) >='+str(rgr)
            cur.execute(trl)
            conn.commit()
            trl= 'create table ' +msk[k]+'_SR'+str(rgr)+' as select '
            trl=trl+msk[k]+'_BioP.* from '+msk[k]+'_BioP inner join tmp
on (tmp.probeset_id = ' +msk[k]+'_BioP.probeset_id) order by '+
msk[k]+'_BioP.probeset_id, '+msk[k]+'_BioP.probe_index'
            cur.execute(trl)
            conn.commit()
```

```
            fp.write('\tTable '), fp.write(msk[k]+'_SR'+str(rgr)),
fp.write(' created.\n')
            pid_tbles.append(msk[k]+'_SR'+str(rgr))
            trl='drop table tmp'
            cur.execute(trl)
            conn.commit()
            cur, conn=UpdateWorkReg(cur, conn, msk[k]+'_SR'+str(rgr),
exp, 'statistical_filter', usr, msk[k]+'_BioP', 'biophysical_filter',
notes)
            k+=1
      onto='select * from '+msk[k-1]+'_SR'+str(rgr)
      cur.execute(onto)
      desc=cur.description
      conn.close()
      fp=UpdateOntology2(usr, pswd, db, desc, exp,
'statistical_filter', '_SR'+str(rgr), notes, str(rgr), fp)
      fp.close()

def DriverFNL(usr, pswd, db, logfile, rgr):
      # first step when combines chip and cel info
      cur, conn= make_connect(usr, pswd, db)
      msk, msk_tab, Lmsk, state=get_mask2(cur)
      #print len(msk), len(tables), len(msk_tab)
      t = datetime.datetime.now()
      EpochSeconds=time.mktime(t.timetuple())
      now = datetime.datetime.fromtimestamp(EpochSeconds)
      fp=open(logfile, 'a')
      fp.write('\n')
      fp.write('Statistical Rigor filters:\n')
      conn.close()
      Fnl_filter(usr, pswd, db, msk, msk_tab, fp, rgr)
```

Appendix C

Sample analysis from Cleansin_Graphics_7_23.py after visual batch inspection (Graphic, Graphic_nobatch, or Graphic_MSR (raw data) functions in same file)

```
def Driver(usr, pswd, db, gfiles, logfile):
      Cel_Probe_Filter(usr, pswd, db, gfiles[0], logfile)
      Cel_Probeset_Filter(usr, pswd, db, gfiles[1], logfile)

def Cel_Probe_Filter(usr, pswd, db, gfile, logfile):
      cur, conn= make_connect(usr, pswd, db)
      msk, Lmsk, state=get_inc_mask(cur)
      cc=zeros(len(Lmsk))
      states=get_unique_states(cur)
      for i in range(len(state)):
            cc[i]=states.index(state[i])
      r.pdf(gfile, height=11, width=8)
      r.par(mfrow=r.c(2,1))
      fp=open(logfile,'a')
      for i in range(len(states)):
            ptr=nonzero(equal(i,cc))
            mu, prbs = zeros(len(ptr), Float), zeros(len(ptr), Float)
            x=range(len(ptr))
            nbr=len(ptr)
            k=0
            for j in ptr:
                  tmp='select signalrawintensity from '+ msk[j]+'_sr4'
                  cur.execute(tmp)
                  rows=cur.fetchall()
                  prbs[k]=len(rows)
                  mu[k]=(sum(rows)[0])/prbs[k]
                  k+=1
            # plot intensities
            r.plot(x, mu, main='Intensity Filter ('+states[i]+')',
xlab='Array Number', ylab='Average Cel Intensity',  pch=21, col='blue',
ylim=r.c(r.mean(mu)-(2.5*r.sd(mu)), r.mean(mu)+(2.5*r.sd(mu))))
            # 2 std dev
            r.lines(x, r.rep(r.mean(mu)-(2*r.sd(mu)), nbr),
col=r.rgb(227/256.,26/256.,28/256.), lty=3, lwd=1)
            r.lines(x, r.rep(r.mean(mu)+(2*r.sd(mu)), nbr),
col=r.rgb(227/256.,26/256.,28/256.), lty=3, lwd=1)
            # 1 std dev
            r.lines(x, r.rep(r.mean(mu)-r.sd(mu), nbr),
col='turquoise', lty=4, lwd=1)
            r.lines(x, r.rep(r.mean(mu)+r.sd(mu), nbr),
col='turquoise', lty=4, lwd=1)
            # mean
            r.lines(x, r.rep(r.mean(mu), nbr), lty=2, lwd=2)
```

```
            #determine outliers
            lowrs1=nonzero(less_equal(mu, ceil(r.mean(mu)-
(2*r.sd(mu)))))
            upprs1=nonzero(greater_equal(mu,
floor(r.mean(mu)+(2*r.sd(mu)))))
            lowrs2=nonzero(less_equal(prbs, ceil(r.mean(prbs)-
(2*r.sd(prbs)))))
            upprs2=nonzero(greater_equal(prbs,
floor(r.mean(prbs)+(2*r.sd(prbs)))))

            for j in lowrs1:
                  cmn=nonzero(equal(j,lowrs2))
                  if len(cmn)==1:
                        r.points(x[j],mu[j], pch='X')
                        tmp="update sample_mask set exclude = true,
description = 'array probe intensities and numbers below 2 stdevs"+"'
where mask_id = '"+msk[ptr[j]]+"'"
                        fp.write('\n'+msk[ptr[j]]+' excluded:\tprobe
intensities and numbers below 2 stdevs')
                  else:
                        cmn=nonzero(equal(j,upprs2))
                        if len(cmn)==1:
                              r.points(x[j],mu[j], pch='X')
                              tmp="update sample_mask set exclude =
true, description = 'array probe intensities below 2 stdevs and numbers
above stdevs"+"' where mask_id = '"+msk[ptr[j]]+"'"
                              fp.write('\n'+msk[ptr[j]]+'
excluded:\tprobe intensities below 2 stdevs and numbers above 2
stdevs')
                        else:
                              r.points(x[j],mu[j], pch='X')
                              tmp="update sample_mask set exclude =
true, description = 'avg array probe intensities below 2 stdevs"+"'
where mask_id = '"+msk[ptr[j]]+"'"
                              fp.write('\n'+msk[ptr[j]]+'
excluded:\tprobe intensities below 2 stdevs')
                  cur.execute(tmp)
                  conn.commit()

            for j in upprs1:
                  cmn=nonzero(equal(j,upprs2))
                  if len(cmn)==1:
                        r.points(x[j],mu[j], pch='X')
                        tmp="update sample_mask set exclude = true,
description = 'array probe intensities and numbers above 2 stdevs"+"'
where mask_id = '"+msk[ptr[j]]+"'"
                        fp.write('\n'+msk[ptr[j]]+' excluded:\tprobe
intensities and numbers above 2 stdevs')
                  else:
                        cmn=nonzero(equal(j,lowrs2))
                        if len(cmn)==1:
                              r.points(x[j],mu[j], pch='X')
```

```
                              tmp="update sample_mask set exclude =
true, description = 'array probe intensities above 2 stdevs and numbers
below stdevs"+"' where mask_id = '"+msk[ptr[j]]+"'"
                              fp.write('\n'+msk[ptr[j]]+'
excluded:\tprobe intensities above 2 stdevs and numbers below 2
stdevs')
                        else:
                              r.points(x[j],mu[j], pch='X')
                              tmp="update sample_mask set exclude =
true, description = 'avg probe numbers below 2 stdevs"+"' where mask_id
= '"+msk[ptr[j]]+"'"
                              fp.write('\n'+msk[ptr[j]]+'
excluded:\tprobe numbers below 2 stdevs')
                  cur.execute(tmp)
                  conn.commit()

            # plot contributing probes
            r.plot(x, prbs, main='Contributing Probe Filter
('+states[i]+')', xlab='Array Number', ylab='Probe Numbers',
ylim=r.c(r.mean(prbs)-(2.5*r.sd(prbs)), r.mean(prbs)+(2.5*r.sd(prbs))),
pch=21, col='blue')

            r.lines(x, r.rep(r.mean(prbs), nbr), lty=2, lwd=2)

            r.lines(x, r.rep(r.mean(prbs)-(2*r.sd(prbs)), nbr),
col=r.rgb(227/256.,26/256.,28/256.), lty=3, lwd=1)
            r.lines(x, r.rep(r.mean(prbs)+(2*r.sd(prbs)), nbr),
col=r.rgb(227/256.,26/256.,28/256.), lty=3, lwd=1)

            r.lines(x, r.rep(r.mean(prbs)-r.sd(prbs), nbr),
col='turquoise', lty=4, lwd=1)
            r.lines(x, r.rep(r.mean(prbs)+r.sd(prbs), nbr),
col='turquoise', lty=4, lwd=1)

            for j in lowrs2:
                  r.points(x[j],prbs[j], pch='X')
                  cmn=nonzero(equal(j,lowrs1))
                  if len(cmn)<1:
                        cmn=nonzero(equal(j,upprs1))
                        if len(cmn)<1:
                              tmp="update sample_mask set exclude =
true, description = 'avg array probe numbers below 2 stdevs"+"' where
mask_id = '"+msk[ptr[j]]+"'"
                              cur.execute(tmp)
                              conn.commit()
                              fp.write('\n'+msk[ptr[j]]+'
excluded:\tprobe numbers below 2 stdevs')

            for j in upprs2:
                  r.points(x[j],prbs[j], pch='X')
                  cmn=nonzero(equal(j,lowrs1))
                  if len(cmn)<1:
                        cmn=nonzero(equal(j,upprs1))
                        if len(cmn)<1:
```

```
                              tmp="update sample_mask set exclude =
true, description = 'avg array probe numbers above 2 stdevs"+"' where
mask_id = '"+msk[ptr[j]]+"'"
                              cur.execute(tmp)
                              conn.commit()
                              fp.write('\n'+msk[ptr[j]]+'
excluded:\tprobe numbers above 2 stdevs')

      conn.close()
      r.dev_off()
      fp.close()



def Cel_Probeset_Filter(usr, pswd, db, gfile, logfile):
      cur, conn= make_connect(usr, pswd, db)
      msk, Lmsk, state=get_inc_mask(cur)
      cc=zeros(len(Lmsk))
      states=get_unique_states(cur)
      for i in range(len(state)):
            cc[i]=states.index(state[i])
      r.pdf(gfile, height=11, width=8)
      r.par(mfrow=r.c(2,1))
      fp=open(logfile,'a')
      for i in range(len(states)):
            ptr=nonzero(equal(i,cc))
            mu, prbs = zeros(len(ptr), Float), zeros(len(ptr), Float)
            x=range(len(ptr))
            nbr=len(ptr)
            k=0
            for j in ptr:
                  # probesets do have a minumum of 4 probes
                  tmp='select distinct(probeset_id) from '+
msk[j]+'_sr4'
                  cur.execute(tmp)
                  rows=cur.fetchall()
                  prbs[k]=len(rows)
                  ps_sgnl=zeros(len(rows), Float)
                  k2=0
                  for l in rows:
                        tmp= 'select signalrawintensity from '+
msk[j]+"_sr4 where probeset_id = '" +l[0]+ "'"
                        cur.execute(tmp)
                        sri=cur.fetchall()
                        ps_sgnl[k2]=(sum(sri)[0])/len(sri)
                        k2+=1
                  mu[k]=r.mean(ps_sgnl)
                  k+=1

            r.plot(x, mu, main='Probeset Intensity Filter
('+states[i]+')', xlab='Array Number', ylab='Average Cel Probeset
Intensity',  pch=21, col='blue', ylim=r.c(r.mean(mu)-(2.5*r.sd(mu)),
r.mean(mu)+(2.5*r.sd(mu))))
```

```
            # 2 std dev
            r.lines(x, r.rep(r.mean(mu)-(2*r.sd(mu)), nbr),
col=r.rgb(227/256.,26/256.,28/256.), lty=3, lwd=1)
            r.lines(x, r.rep(r.mean(mu)+(2*r.sd(mu)), nbr),
col=r.rgb(227/256.,26/256.,28/256.), lty=3, lwd=1)
            # 1 std dev
            r.lines(x, r.rep(r.mean(mu)-r.sd(mu), nbr),
col='turquoise', lty=4, lwd=1)
            r.lines(x, r.rep(r.mean(mu)+r.sd(mu), nbr),
col='turquoise', lty=4, lwd=1)
            # mean
            r.lines(x, r.rep(r.mean(mu), nbr), lty=2, lwd=2)

            #determine outliers
            lowrs1=nonzero(less_equal(mu, ceil(r.mean(mu)-
(2*r.sd(mu)))))
            upprs1=nonzero(greater_equal(mu,
floor(r.mean(mu)+(2*r.sd(mu)))))
            lowrs2=nonzero(less_equal(prbs, ceil(r.mean(prbs)-
(1.5*r.sd(prbs)))))

            for j in lowrs1:
                  cmn=nonzero(equal(j,lowrs2))
                  if len(cmn)==1:
                        r.points(x[j],mu[j], pch='X')
                        tmp="update sample_mask set exclude = true,
description = 'array probeset intensities and numbers below 2(1.5)
stdevs"+"' where mask_id = '"+msk[ptr[j]]+"'"
                        fp.write('\n'+msk[ptr[j]]+' excluded:\tprobeset
intensities and numbers below 2(1.5) stdevs')
                  else:
                        r.points(x[j],mu[j], pch='X')
                        tmp="update sample_mask set exclude = true,
description = 'avg array probeset intensities below 2 stdevs"+"' where
mask_id = '"+msk[ptr[j]]+"'"
                        fp.write('\n'+msk[ptr[j]]+' excluded:\tprobeset
intensities below 2 stdevs')
                  cur.execute(tmp)
                  conn.commit()

            for j in upprs1:
                  cmn=nonzero(equal(j,lowrs2))
                  if len(cmn)==1:
                        r.points(x[j],mu[j], pch='X')
                        tmp="update sample_mask set exclude = true,
description = 'array probeset intensities above 2 stdevs and probesets
below 1.5 stdevs"+"' where mask_id = '"+msk[ptr[j]]+"'"
                        fp.write('\n'+msk[ptr[j]]+' excluded:\tprobeset
intensities above 2 stdevs and probesets below 1.5 stdevs')
                  else:
                        r.points(x[j],mu[j], pch='X')
                        tmp="update sample_mask set exclude = true,
description = 'avg probeset intensities above 2 stdevs"+"' where
mask_id = '"+msk[ptr[j]]+"'"
```

```
                        fp.write('\n'+msk[ptr[j]]+' excluded:\tprobeset
numbers below 1.5 stdevs')
                  cur.execute(tmp)
                  conn.commit()

            # plot contributing probes
            r.plot(x, prbs, main='Contributing Probeset Filter
('+states[i]+')', xlab='Array Number', ylab='Probe Numbers',
ylim=r.c(r.mean(prbs)-(2.5*r.sd(prbs)), r.mean(prbs)+(2.5*r.sd(prbs))),
pch=21, col='blue')

            # mean
            r.lines(x, r.rep(r.mean(prbs), nbr), lty=2, lwd=2)

            #  2 std dev
            r.lines(x, r.rep(r.mean(prbs)-(2*r.sd(prbs)), nbr),
col=r.rgb(227/256.,26/256.,28/256.), lty=3, lwd=1)
            r.lines(x, r.rep(r.mean(prbs)+(2*r.sd(prbs)), nbr),
col=r.rgb(227/256.,26/256.,28/256.), lty=3, lwd=1)

            # 1 std dev
            r.lines(x, r.rep(r.mean(prbs)-r.sd(prbs), nbr),
col='turquoise', lty=4, lwd=1)
            r.lines(x, r.rep(r.mean(prbs)+r.sd(prbs), nbr),
col='turquoise', lty=4, lwd=1)

            for j in lowrs2:
                  r.points(x[j],prbs[j], pch='X')
                  cmn=nonzero(equal(j,lowrs1))
                  if len(cmn)<1:
                        tmp="update sample_mask set exclude = true,
description = 'array probeset numbers below 1.5 stdevs"+"' where
mask_id = '"+msk[ptr[j]]+"'"
                        cur.execute(tmp)
                        conn.commit()
                        fp.write('\n'+msk[ptr[j]]+' excluded:\tprobeset
numbers below 1.5 stdevs')

      conn.close()
      r.dev_off()
      fp.close()
```

Appendix D

Aggregate BaFL cleansed samples from the Aggregation.py file

Individuals within the same disease class, if the sample size is large the population is randomized and divided into sub-populations before intersecting the entire population, for computational efficiency.

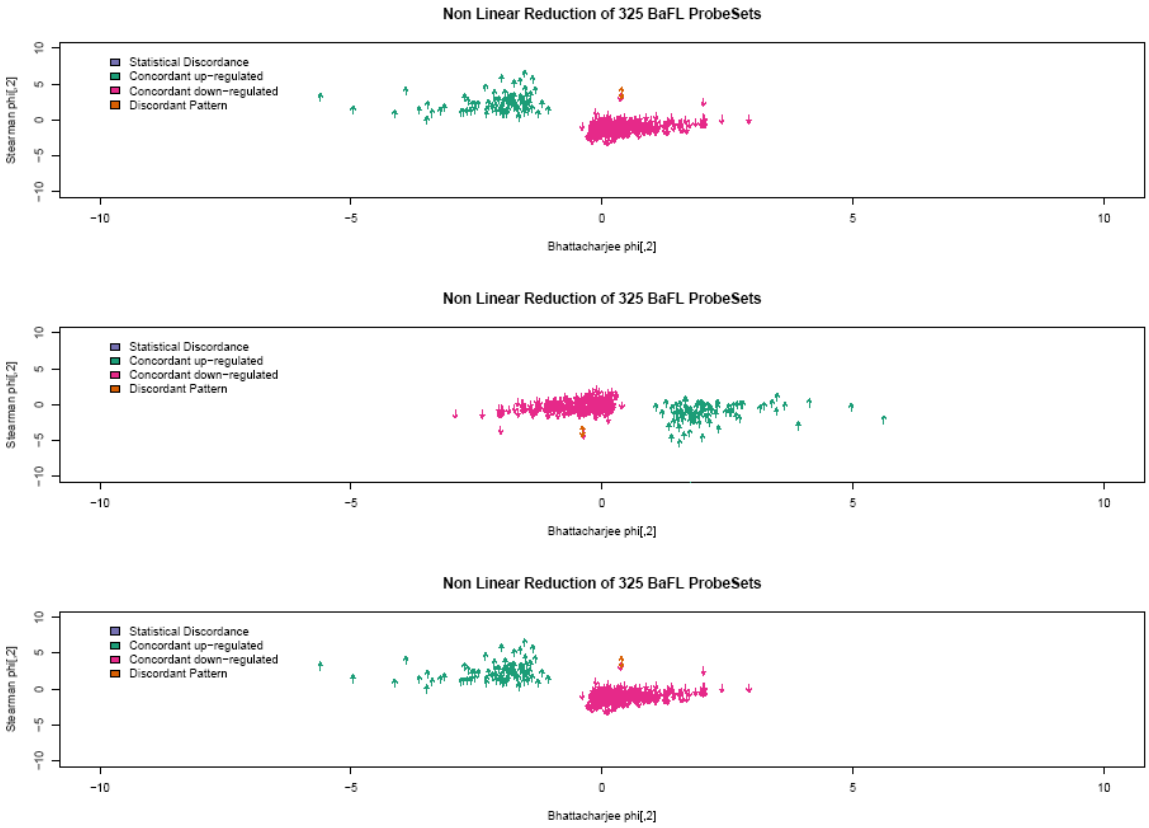
```
def StateIntersect(usr, pswd, db, logfile, rgr=4):
      cur, conn= make_connect(usr, pswd, db)
      states=get_unique_states(cur)
      exp=get_exp(usr, pswd, db)
      notes='intersection of '
      fp=open(logfile, 'a')
      for i in states:
            k=i
            pt=find(i,' ')
            if pt>0:
                  k=i[:pt]+i[pt+1:]
            msk, Lmsk, state, cel=get_inc_state(cur, i)
            if len(msk)>50:
                  div=2
                  while len(msk)/div>30:
                        div+=1
                  cur, conn =splitIntersect(cur, conn, k, msk, div)
            else:
                  tmp= 'create table Intersect_'+k+' as select '
+msk[0]+'_sr'+str(rgr)+'.probeset_id, '
+msk[0]+'_sr'+str(rgr)+'.pm_mm_other, '+msk[0]+'_sr'+str(rgr)+'.x,
'+msk[0]+'_sr'+str(rgr)+'.y, '+msk[0]+'_sr'+str(rgr)+'.probe_index, '
                  tmp= tmp+ msk[0]+'_sr'+str(rgr)+'.signalrawintensity
as '+msk[0]
                  for j in range(1, len(msk)):
                        tmp= tmp+ ',
'+msk[j]+'_sr'+str(rgr)+'.signalrawintensity as '+msk[j]
                  tmp= tmp +' from ' +msk[0]+'_sr'+str(rgr)+' inner
join '+msk[1]+'_sr'+str(rgr)+' on ('+msk[0]+'_sr'+str(rgr)+'.x =
'+msk[1]+'_sr'+str(rgr)+'.x and '+msk[0]+'_sr'+str(rgr)+'.y =
'+msk[1]+'_sr'+str(rgr)+'.y ) '
                  for j in range(2, len(msk)):
                        tmp=tmp+ 'inner join '+msk[j]+'_sr'+str(rgr)+'
on (' +msk[j]+'_sr'+str(rgr)+'.x = '+msk[j-1]+'_sr'+str(rgr)+'.x and
'+msk[j]+'_sr'+str(rgr)+'.y = '+msk[j-1]+'_sr'+str(rgr)+'.y ) '
                  tmp=tmp +' order by
'+msk[0]+'_sr'+str(rgr)+'.probeset_id,
'+msk[0]+'_sr'+str(rgr)+'.probe_index'
                  cur.execute(tmp)
                  conn.commit()
```

```
            cur, conn=UpdateWorkReg(cur, conn, 'Intersect_'+k, exp,
'Intersect_'+k, usr, k+'_sr'+str(rgr), 'statistical_filter', notes+i)

      desc=['probeset_id','pm_mm_other','x','y','probe_index','SRIs']
            fp.write('\nCreated Table:\tIntersect_'+k+'\n')
            fp=UpdateOntology3(usr, pswd, db, desc, exp,
'Intersect_'+k, 'Intersect_'+k, notes+i, 'statistical rigor not
enforced', fp)
      conn.close()
      fp.close()



# Intersect all two class datasets

def ModelIntersect(usr, pswd, db, logfile, gt=4):
      cur, conn= make_connect(usr, pswd, db)
      states=get_unique_states(cur)
      vals=['probe_index', 'x', 'y', 'pm_mm_other']
      exp=get_exp(usr, pswd, db)
      notes='intersection of '
      fp=open(logfile, 'a')
      for i in range(len(states)-1):
            class1=states[i]
            pt=find(class1,' ')
            if pt>0:
                  class1=states[i][:pt]+states[i][pt+1:]
            msk1, Lmsk1, state1, cel1=get_inc_state(cur, states[i])
            tmp1='create table temp1 as select
intersect_'+class1+'.probeset_id'
            for k in vals:
                  tmp1=tmp1+ ', intersect_'+class1+'.'+k
            for k in msk1:
                  tmp1=tmp1+ ', intersect_'+class1+'.'+k+' as '+k
            for j in range(i+1, len(states)):
                  tmp=tmp1
                  class2=states[j]
                  pt=find(class2,' ')
                  if pt>0:
                        class2=states[j][:pt]+states[j][pt+1:]
                  msk2, Lmsk2, state2, cel2=get_inc_state(cur,
states[j])
                  for k in msk2:
                        tmp=tmp+ ', intersect_'+class2+'.'+k+' as '+k
                  tmp=tmp+' where intersect_'+class1+'.x =
intersect_'+class2+'.x and intersect_'+class1+'.y =
intersect_'+class2+'.y '
                  tmp=tmp+ 'order by intersect_'+class1+'.probeset_id,
intersect_'+class1+'.probe_index'
                  cur.execute(tmp)
                  conn.commit()
                  #print '\n\n', tmp, '\n\n'
```

```
                  tmp= 'create table temp2 as select count(*),
probeset_id from temp1 group by probeset_id having count(*) >=
'+str(gt)
                  cur.execute(tmp)
                  conn.commit()
                  tmp='create table '+class1+'_'+class2+'_'+str(gt)+'
as select temp1.* from temp1, temp2 where
temp1.probeset_id=temp2.probeset_id order by temp1.probeset_id,
temp1.probe_index'
                  cur.execute(tmp)
                  conn.commit()
                  tmp='drop table temp1'
                  cur.execute(tmp)
                  conn.commit()
                  tmp='drop table temp2'
                  cur.execute(tmp)
                  conn.commit()
                  cur, conn=UpdateWorkReg(cur, conn,
class1+'_'+class2+'_'+str(gt), exp, class1+'_'+class2+'_'+str(gt), usr,
'Intersect_'+class1+', Intersect_'+class2, 'Intersect_'+class1+',
Intersect_'+class2, notes+class1+' and '+class2)
                  desc=vals[:]
                  desc.append('SRIs')
                  fp.write('\nCreated
Table:\t'+class1+'_'+class2+'_'+str(gt)+'\n')
                  fp=UpdateOntology3(usr, pswd, db, desc, exp,
class1+'_'+class2+'_'+str(gt), class1+'_'+class2+'_'+str(gt),
notes+class1+' and '+class2, 'statistical rigor of '+str(gt)+'
enforced', fp)
      conn.close()
      fp.close()
```

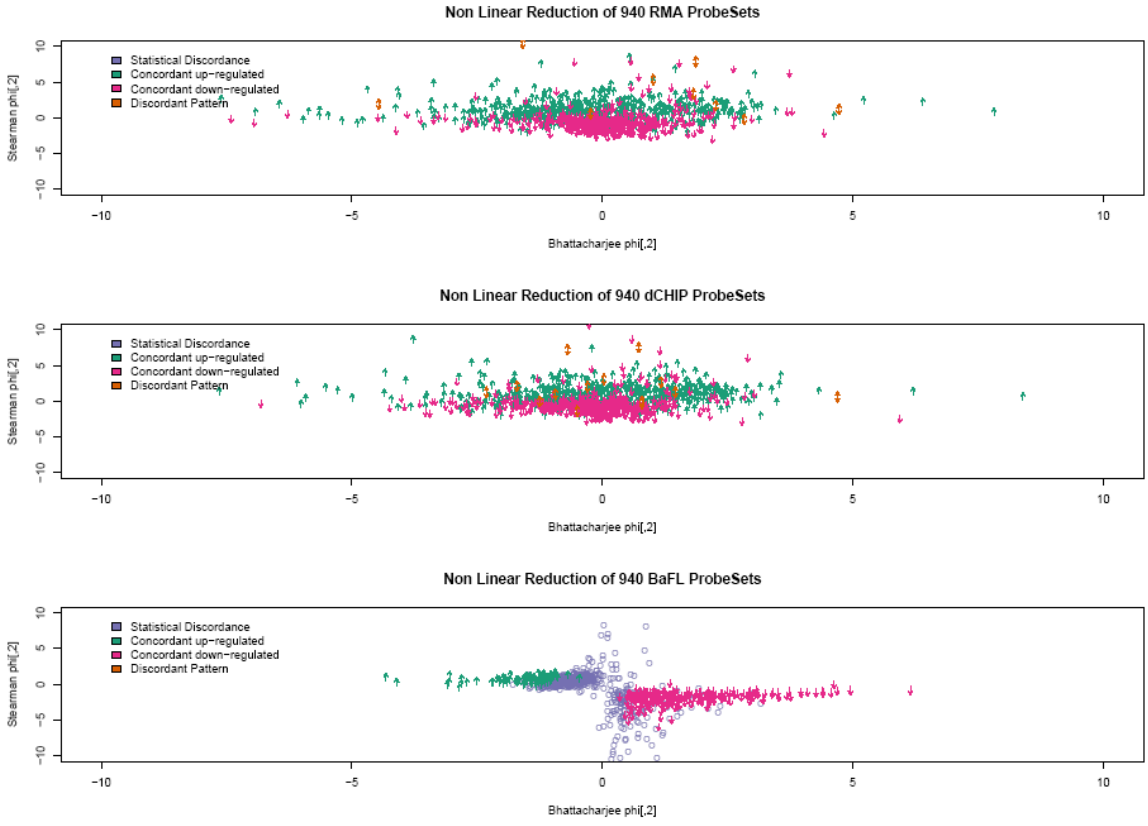
Appendix E

The result of permuting the gene order does not disrupt the latent structure observed across the Bhaattacharjee BaFL dataset and the Stearman BaFL dataset, for the intersecting 325 ProbeSets assessed to be differentially expressed ($\alpha = 0.05$).



Appendix F

The 940 ProbeSets which were retained through the BaFL pipeline and an agreement by RMA and dCHIP to be differentially expressed ($\alpha = 0.05$). The latent structure is still observed with the BaFL interpretation of the ProbeSets.



BIBLIOGRAPHY

BIBLIOGRAPHY

Chapter 1:

1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al*: **Minimum information about a Microarray experiment (MIAME)-toward standards for Microarray data**. *Nature genetics* 2001, **29**(4):365-371.

2. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays**. *Nature genetics* 1999, **21**(1 Suppl):20-24.

3. Fixman M, Freire JJ: **Theory of DNA melting curves**. *Biopolymers* 1977, **16**(12):2693-2704.

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

5. Southern EM: **DNA Microarrays. History and overview**. In: *Methods in Molecular Biology.* Edited by Rampal JB, vol. 170: Humana Press; 2001: 15.

6. Southern EM: **DNA Microarrays**. In: *DNA Arrays Methods and Protocols.* Edited by Rampal JB, vol. 170. Totoja, NJ: Humana Press: 1-15.

7. Yalow RS, Berson SA: **Immunoassay of endogenous plasma insulin in man**. *J Clin Invest* 1960, **39**:1157-1175.

8. **Affymetrix.com** [http://www.affymetrix.com]

9. **GeneChip® Expression Analysis Technical Manual** [http://www.affymetrix.com/support/technical/manual/expression_manual.affx]

10. Southern E, Mir K, Shchepinov M: **Molecular interactions on Microarrays**. *Nature genetics* 1999, **21**(1 Suppl):5-9.

11. **Nimblegen.com** [http://www.nimblegen.com/]

12. **Agilent.com** [http://www.agilent.com]

13. Bowtell DD: **Options available--from start to finish--for obtaining expression data by Microarray**. *Nature genetics* 1999, **21**(1 Suppl):25-32.

14. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G: **Making and reading Microarrays**. *Nature genetics* 1999, **21**(1 Suppl):15-19.

15. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA Microarrays**. *Nature genetics* 1999, **21**(1 Suppl):10-14.

16. Brown PO, Botstein D: **Exploring the new world of the genome with DNA Microarrays**. *Nature genetics* 1999, **21**(1 Suppl):33-37.

17. Flikka K, Yadetie F, Laegreid A, Jonassen I: **XHM: a system for detection of potential cross hybridizations in DNA Microarrays**. *BMC Bioinformatics* 2004, **5**:117.

18. Kumari S, Verma LK, Weller JW: **AffyMAPSDetector: a software tool to characterize Affymetrix GeneChip expression arrays with respect to SNPs**. *BMC Bioinformatics* 2007, **8**:276.

19. Seman Kachalo ZAaJL: **Method of Microarray Data Analysis III**. *Paper from Camda '02* 2002:185-199.

20. Wren JD, Kulkarni A, Joslin J, Butow RA, Garner HR: **Cross-hybridization on PCR-spotted Microarrays**. *IEEE Eng Med Biol Mag* 2002, **21**(2):71-75.

21. Bevilacqua PC, SantaLucia J, Jr.: **The biophysics of RNA**. *ACS Chem Biol* 2007, **2**(7):440-4

22. SantaLucia J, Jr., Hicks D: **The thermodynamics of DNA structural motifs**. *Annu Rev Biophys Biomol Struct* 2004, **33**:415-440.

23. Watkins NE, Jr., SantaLucia J, Jr.: **Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes**. *Nucleic Acids Res* 2005, **33**(19):6258-6267.

24. Weckx S, Carlon E, DeVuyst L, Van Hummelen P: **Thermodynamic behavior of short oligonucleotides in Microarray hybridizations can be described using Gibbs free energy in a nearest-neighbor model**. *J Phys Chem B* 2007, **111**(48):13583-13590.

25. SantaLucia J, Jr., Allawi HT, Seneviratne PA: **Improved nearest-neighbor parameters for predicting DNA duplex stability**. *Biochemistry* 1996, **35**(11):3555-3562.

26. SantaLucia J, Jr., Turner DH: **Measuring the thermodynamics of RNA secondary structure formation**. *Biopolymers* 1997, **44**(3):309-319.

27. Wu P, Nakano S, Sugimoto N: **Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation**. *Eur J Biochem* 2002, **269**(12):2821-2830.

28. Allawi HT, SantaLucia J, Jr.: **Thermodynamics of internal C.T mismatches in DNA**. *Nucleic Acids Res* 1998, **26**(11):2694-2701.

29. Shchepinov MS, Case-Green SC, Southern EM: **Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays**. *Nucleic Acids Res* 1997, **25**(6):1155-1161.

30. Rouillard JM, Zuker M, Gulari E: **OligoArray 2.0: design of oligonucleotide probes for DNA Microarrays using a thermodynamic approach**. *Nucleic Acids Res* 2003, **31**(12):3057-3062.

31. Chakravarti A: **Population genetics--making sense out of sequence**. *Nature genetics* 1999, **21**(1 Suppl):56-60.

32. Venables JP: **Aberrant and alternative splicing in cancer**. *Cancer Res* 2004, **64**(21):7647-7654.

33. Venables JP: **Unbalanced alternative splicing and its significance in cancer**. *Bioessays* 2006, **28**(4):378-386.

34. **Gene Probes: A Practical Approach, Vols. 1 & 2**: Oxford University Press; 1995.

35. Draghici S, Khatri P, Eklund AC, Szallasi Z: **Reliability and reproducibility issues in DNA Microarray measurements**. *Trends Genet* 2006, **22**(2):101-109.

36. Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, Yoneyama M, Sasaki M: **Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes**. *Biochemistry* 1995, **34**(35):11211-11216.

37. Straus NA, Bonner TI: **Temperature dependence of RNA-DNA hybridization kinetics**. *Biochim Biophys Acta* 1972, **277**(1):87-95.

38. Bengtsson H, Jonsson G, Vallon-Christersson J: **Calibration and assessment of channel-specific biases in Microarray data with extended dynamical range**. *BMC Bioinformatics* 2004, **5**:177.

39. Shi L, Tong W, Su Z, Han T, Han J, Puri RK, Fang H, Frueh FW, Goodsaid FM, Guo L *et al*: **Microarray scanner calibration curves: characteristics and implications**. *BMC Bioinformatics* 2005, **6 Suppl 2**:S11.

40. Borden JR, Paredes CJ, Papoutsakis ET: **Diffusion, mixing, and associated dye effects in DNA-Microarray hybridizations**. *Biophys J* 2005, **89**(5):3277-3284.

41. Naderi A, Ahmed AA, Wang Y, Brenton JD, Caldas C: **Optimal amounts of fluorescent dye improve expression Microarray results in tumor specimens**. *Mol Biotechnol* 2005, **30**(2):151-154.

42. Uchida S, Nishida Y, Satou K, Muta S, Tashiro K, Kuhara S: **Detection and normalization of biases present in spotted cDNA Microarray data: a composite**

method addressing dye, intensity-dependent, spatially-dependent, and print-order biases. *DNA Res* 2005, **12**(1):1-7.

43. Cuff JA, Coates GM, Cutts TJ, Rae M: **The Ensembl computing architecture**. *Genome Res* 2004, **14**(5):971-975.

44. Shields R: **MIAME, we have a problem**. *Trends Genet* 2006, **22**(2):65-66.

45. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data**. *Bioinformatics (Oxford, England)* 2006, **22**(8):943-949.

46. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.

47. Irizarry RA, Wu Z, Jaffee HA: **Comparison of Affymetrix GeneChip expression measures**. *Bioinformatics (Oxford, England)* 2006, **22**(7):789-794.

48. Li C, Hung Wong W: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. *Genome biology* 2001, **2**(8):RESEARCH0032.

49. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection**. *Proc Natl Acad Sci U S A* 2001, **98**(1):31-36.

50. Meh DA, Mosesson MW, Siebenlist KR, Simpson-Haidaris PJ, Brennan SO, DiOrio JP, Thompson K, Di Minno G: **Fibrinogen naples I (B beta A68T) nonsubstrate thrombin-binding capacities**. *Thromb Res* 2001, **103**(1):63-73.

51. Bassett DE, Jr., Eisen MB, Boguski MS: **Gene expression informatics--it's all in your mine**. *Nature genetics* 1999, **21**(1 Suppl):51-55.

52. G. Parmigiani ESG, R. A. Irizarry, S. L. Zeger: **The Analysis of Gene Expression Data**. New York: Springer; 2003.

53. Ratushna VG, Weller JW, Gibas CJ: **Secondary structure in the target as a confounding factor in synthetic oligomer Microarray design**. *BMC Genomics* 2005, **6**(1):31.

54. SantaLucia J, Jr., Kierzek R, Turner DH: **Effects of GA mismatches on the structure and thermodynamics of RNA internal loops**. *Biochemistry* 1990, **29**(37):8813-8819.

55. Sugimoto N, Nakano M, Nakano S: **Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes**. *Biochemistry* 2000, **39**(37):11270-11281.

56. SantaLucia J, Jr., Kierzek R, Turner DH: **Stabilities of consecutive A.C, C.C, G.G, U.C, and U.U mismatches in RNA internal loops: Evidence for stable hydrogen-bonded U.U and C.C.+ pairs**. *Biochemistry* 1991, **30**(33):8242-8251.

57. Ioannidis JP: **Microarrays and molecular research: noise discovery?** *Lancet* 2005, **365**(9458):454-455.

58. Cole KA, Krizman DB, Emmert-Buck MR: **The genetics of cancer--a 3D model**. *Nature genetics* 1999, **21**(1 Suppl):38-41.

59. Kachalo S. AZ, and Liang J.: **Assessing the potential effect of cross-hybridization on oligonucleotide Microarrays**. In: *Methods of Microarray Data Analysis III.* Edited by Kimberly F. Johnson SML. Norwell: Kluwer Academic Publishers; 2003.

60. Lander ES: **Array of hope**. *Nature genetics* 1999, **21**(1 Suppl):3-4.

61. [http://www.abrf.org/]

62. Hacia JG: **Resequencing and mutational analysis using oligonucleotide Microarrays**. *Nature genetics* 1999, **21**(1 Suppl):42-47.

63. Brown VM, Ossadtchi A, Khan AH, Cherry SR, Leahy RM, Smith DJ: **High-throughput imaging of brain gene expression**. *Genome Res* 2002, **12**(2):244-254.

64. **Netaffx Analysis Center** [https://www.affymetrix.com/site/login/login.affx]

65. Carlon E, Heim T, Wolterink JK, Barkema GT: **Comment on "Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays"**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006, **73**(6 Pt 1):063901; author reply 063902.

66. Lincoln SE: **CHI Total Microarray Data Analysis and Interpretation**. In. Edited by Panel. Washington D.C.; 2006: Stated while addressing the morning panel.

67. Ho J, Hwang WL: **Automatic Microarray spot segmentation using a Snake-Fisher model**. *IEEE Trans Med Imaging* 2008, **27**(6):847-857.

68. Shields R: **The emperor's new clothes revisited**. *Trends Genet* 2006, **22**(9):463.

69. Fridlyand SDaJ: **Introduction to Classification in Microarray Experiments**. In: *DNA Arrays Methods and Protocols.* Edited by Rampal JB, vol. 170. Totoja, NJ: Humana Press: 132-149.

70. Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF: **GEMS: a system for automated cancer diagnosis and biomarker discovery from Microarray gene expression data**. *International journal of medical informatics* 2005, **74**(7-8):491-503.

71. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S: **A comprehensive evaluation of multicategory classification methods for Microarray gene expression cancer diagnosis**. *Bioinformatics (Oxford, England)* 2005, **21**(5):631-643.

72. Ringner M, Peterson C, Khan J: **Analyzing array data using supervised methods**. *Pharmacogenomics* 2002, **3**(3):403-415.

73. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *Proc Natl Acad Sci U S A* 2000, **97**(18):10101-10106.

74. Brody JP, Williams BA, Wold BJ, Quake SR: **Significance and statistical errors in the analysis of DNA Microarray data**. *Proc Natl Acad Sci U S A* 2002, **99**(20):12975-12978.

75. Fathallah-Shaykh HM: **Microarrays: applications and pitfalls**. *Arch Neurol* 2005, **62**(11):1669-1672.

76. van Steensel B: **Mapping of genetic and epigenetic regulatory networks using Microarrays**. *Nature genetics* 2005, **37 Suppl**:S18-24.

77. Alpaydin E: **Introduction to Machine Learning**. Cambridge: The MIT Press; 2004.

78. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M *et al*: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proc Natl Acad Sci U S A* 2001, **98**(24):13790-13795.

79. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**(5439):531-537.

80. Li L, Weinberg CR, Darden TA, Pedersen LG: **Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method**. *Bioinformatics (Oxford, England)* 2001, **17**(12):1131-1142.

81. Ntzani EE, Ioannidis JP: **Predictive ability of DNA Microarrays for cancer outcomes and correlates: an empirical assessment**. *Lancet* 2003, **362**(9394):1439-1444.

82. Ringner M, Peterson C: **Microarray-based cancer diagnosis with artificial neural networks**. *Biotechniques* 2003, **Suppl**:30-35.

83. Statnikov A, Wang L, Aliferis CF: **A comprehensive comparison of random forests and support vector machines for Microarray-based cancer classification**. *BMC Bioinformatics* 2008, **9**:319.

84. Manly BFJ: **Multivariate Statistical Methods**, 3rd edn. Washington D.C.: Chapman & Hall/CRC; 2005.

85. Edward Keedwell AN: **Intelligent Bioinformatics**
The application of artificial intelligence techniques to bioinformatics problems. Hoboken: John Wiley & Sons; 2005.

86. Rosner B: **Fundamentals of Biostatistics**, 5th edn. Pacific Grove: Duxbury; 2000.

87. Davis JG, M.: **The Relationship between Precision-Recall and ROC Curves**. *ICML Proceedings of the 23rd International Conference on Machine Learning* 2006:8.

88. Ian H. Witten EF: **Data Mining: Practical Machine Learning Tools and Techniques**, 2nd edn: Morgan Kaufmann; 2005.

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

90. Minna JD, Roth JA, Gazdar AF: **Focus on lung cancer**. *Cancer Cell* 2002, **1**(1):49-52.

Chapter 2:

1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al*: **Minimum information about a Microarray experiment (MIAME)-toward standards for Microarray data**. *Nature genetics* 2001, **29**(4):365-371.

2. Fridlyand SDaJ: **Introduction to Classification in Microarray Experiments**. In: *DNA Arrays Methods and Protocols.* Edited by Rampal JB, vol. 170. Totoja, NJ: Humana Press: 132-149.

3. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays**. *Nature genetics* 1999, **21**(1 Suppl):20-24.

4. Southern EM: **DNA Microarrays**. In: *DNA Arrays Methods and Protocols.* Edited by Rampal JB, vol. 170. Totoja, NJ: Humana Press: 1-15.

5. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.

6. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G *et al*: **Multiple-laboratory comparison of Microarray platforms**. *Nat Methods* 2005, **2**(5):345-350.

7. Shields R: **MIAME, we have a problem**. *Trends Genet* 2006, **22**(2):65-66.

8. Shields R: **The emperor's new clothes revisited**. *Trends Genet* 2006, **22**(9):463.
9. Flikka K, Yadetie F, Laegreid A, Jonassen I: **XHM: a system for detection of potential cross hybridizations in DNA Microarrays**. *BMC Bioinformatics* 2004, **5**:117.

10. Wren JD, Kulkarni A, Joslin J, Butow RA, Garner HR: **Cross-hybridization on PCR-spotted Microarrays**. *IEEE Eng Med Biol Mag* 2002, **21**(2):71-75.

11. Kumari S, Verma LK, Weller JW: **AffyMAPSDetector: a software tool to characterize Affymetrix GeneChip expression arrays with respect to SNPs**. *BMC Bioinformatics* 2007, **8**:276.

12. Ratushna VG, Weller JW, Gibas CJ: **Secondary structure in the target as a confounding factor in synthetic oligomer Microarray design**. *BMC Genomics* 2005, **6**(1):31.

13. Deshmukh H: **Modeling the Physical Parameters Affecting the Measurements from Microarrays**. Fairfax: George Mason University; 2006.

14. Bengtsson H, Jonsson G, Vallon-Christersson J: **Calibration and assessment of channel-specific biases in Microarray data with extended dynamical range**. *BMC Bioinformatics* 2004, **5**:177.

15. Kachalo S. AZ, and Liang J.: **Assessing the potential effect of cross-hybridization on oligonucleotide Microarrays**. In: *Methods of Microarray Data Analysis III.* Edited by Kimberly F. Johnson SML. Norwell: Kluwer Academic Publishers; 2003.

16. Shi L, Tong W, Su Z, Han T, Han J, Puri RK, Fang H, Frueh FW, Goodsaid FM, Guo L *et al*: **Microarray scanner calibration curves: characteristics and implications**. *BMC Bioinformatics* 2005, **6 Suppl 2**:S11.

17. Howard BH: **Control of Variability**. *Institute for Laboratory Animal Research* 2002, **43**(4):7.

18. Yalow RS, Berson SA: **Immunoassay of endogenous plasma insulin in man**. *J Clin Invest* 1960, **39**:1157-1175.

19. Irizarry RA: **affy**. In.: Bioconductor.

20. Draghici S, Khatri P, Eklund AC, Szallasi Z: **Reliability and reproducibility issues in DNA Microarray measurements**. *Trends Genet* 2006, **22**(2):101-109.

21. Miron M, Nadon R: **Inferential literacy for experimental high-throughput biology**. *Trends Genet* 2006, **22**(2):84-89.

22. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data**. *Bioinformatics (Oxford, England)* 2006, **22**(8):943-949.

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

24. Li C, Hung Wong W: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. *Genome biology* 2001, **2**(8):RESEARCH0032.

25. Binder H, Preibisch S: **Specific and nonspecific hybridization of oligonucleotide probes on Microarrays**. *Biophys J* 2005, **89**(1):337-352.

26. Carlon E, Heim T, Wolterink JK, Barkema GT: **Comment on "Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays"**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006, **73**(6 Pt 1):063901; author reply 063902.

27. Ferrantini A, Carlon E: **On the relationship between perfect matches and mismatches in Affymetrix Genechips**. *Gene* 2008, **422**(1-2):1-6.

28. SantaLucia J, Jr., Kierzek R, Turner DH: **Stabilities of consecutive A.C, C.C, G.G, U.C, and U.U mismatches in RNA internal loops: Evidence for stable hydrogen-bonded U.U and C.C.+ pairs**. *Biochemistry* 1991, **30**(33):8242-8251.

29. Shchepinov MS, Case-Green SC, Southern EM: **Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays**. *Nucleic Acids Res* 1997, **25**(6):1155-1161.

30. Sugimoto N, Nakano M, Nakano S: **Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes**. *Biochemistry* 2000, **39**(37):11270-11281.

31. Futschik ME, Reeve A, Kasabov N: **Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue**. *Artif Intell Med* 2003, **28**(2):165-189.

32. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**(5439):531-537.

33. Lu Y, Lemon W, Liu PY, Yi Y, Morrison C, Yang P, Sun Z, Szoke J, Gerald WL, Watson M *et al*: **A gene expression signature predicts survival of patients with stage I non-small cell lung cancer**. *PLoS medicine* 2006, **3**(12):e467.

34. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S: **A comprehensive evaluation of multicategory classification methods for Microarray gene expression cancer diagnosis**. *Bioinformatics* 2005, **21**(5):631-643.

35. Berrar DP, Downes CS, Dubitzky W: **Multiclass cancer classification using gene expression profiling and probabilistic neural networks**. *Pac Symp Biocomput* 2003:5-16.

36. **GeneChip® Expression Analysis Technical Manual** [http://www.affymetrix.com/support/technical/manual/expression_manual.affx]

37. Seman Kachalo ZAaJL: **Method of Microarray Data Analysis III**. *Paper from Camda '02* 2002:185-199.

38. Michael Stonebraker LAR, Michael Hirohama **The Design of POSTGRES**. In., 8.0.3 edn: IEEE Transactions on Knowledge and Data Engineering 1986.

39. Rossum Gv: **Python**. In.

40. Gregorio FD: **psycopg2**. In., 2 2.0.2 edn: Psycopg is a PostgreSQL database adapter for the Python programming language. Its main advantages are that it supports the full Python DBAPI 2.0 and it is thread safe at level 2. It was designed for heavily multi-threaded applications that create and destroy lots of cursors and make a conspicuous number of concurrent INSERTs or UPDATEs. The psycopg distribution includes ZPsycopgDA, a Zope Database Adapter.

41. R DCT: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing.

42. Walter Moreira GW: **rpy**. In., 1.0 edn: RPy is a very simple, yet robust, Python interface to the R Programming Language. It can manage all kinds of R objects and can execute arbitrary R functions (including the graphic functions). All errors from the R language are converted to Python exceptions. Any module installed for the R system can be used from within Python.

43. Rouillard JM, Zuker M, Gulari E: **OligoArray 2.0: design of oligonucleotide probes for DNA Microarrays using a thermodynamic approach**. *Nucleic Acids Res* 2003, **31**(12):3057-3062.

44. Andy Dustman JEaMT: **MySQLdb**. In., 1.2.0 edn: MySQL support for Python. MySQL versions 3.23-25.21; and Python versions 22.23-22.25 are supported. MySQLdb is the Python DB API-22.20 interface. _mysql is a low-level API similiar to the MySQL C API. ZMySQLDA is a Database Adapter for Zope22.

45. Cuff JA, Coates GM, Cutts TJ, Rae M: **The Ensembl computing architecture**. *Genome Res* 2004, **14**(5):971-975.

46. **Affymetrix.com** [http://www.affymetrix.com]

47. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M *et al*: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proc Natl Acad Sci U S A* 2001, **98**(24):13790-13795.

48. Bhattacharjee Arindham WGR, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, Massimo Loda, Griffin Weber, Eugene J. Mark, Eric S. Lander, Wing Wong, Bruce E. Johnson, Todd R. Golub, David J. Sugarbaker, and Mathew Meyerson: **Classificaton of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *PNAS* 2001, **98**(24):13790-13795.

49. Stearman RS, Dwyer-Nield L, Zerbe L, Blaine SA, Chan Z, Bunn PA, Jr., Johnson GL, Hirsch FR, Merrick DT, Franklin WA *et al*: **Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model**. *Am J Pathol* 2005, **167**(6):1763-1775.

50. Ioannidis JP: **Microarrays and molecular research: noise discovery?** *Lancet* 2005, **365**(9458):454-455.

51. Rosner B: **Fundamentals of Biostatistics**, 5th edn. Pacific Grove: Duxbury; 2000.

52. Alpaydin E: **Introduction to Machine Learning**. Cambridge: The MIT Press; 2004.

53. **Linear Discriminant Analysis, A Brief tutorial** [http://www.music.mcgill.ca/~ich/classes/mumt611/classifiers/lda_theory.pdf]

54. **Random Forest** [http://oz.berkeley.edu/users/breiman/randomforest2001.pdf]

55. **Looking Inside the Black Box** [http://stat-www.berkeley.edu/users/breiman/wald2002-2.pdf]

56. Diaz-Uriarte R, Alvarez de Andres S: **Gene selection and classification of Microarray data using random forest**. *BMC Bioinformatics* 2006, **7**:3.

57. Manly BFJ: **Multivariate Statistical Methods**, 3rd edn. Washington D.C.: Chapman & Hall/CRC; 2005.

58. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection**. *Proc Natl Acad Sci U S A* 2001, **98**(1):31-36.

59. Borden JR, Paredes CJ, Papoutsakis ET: **Diffusion, mixing, and associated dye effects in DNA-Microarray hybridizations**. *Biophys J* 2005, **89**(5):3277-3284.

60. Uchida S, Nishida Y, Satou K, Muta S, Tashiro K, Kuhara S: **Detection and normalization of biases present in spotted cDNA Microarray data: a composite method addressing dye, intensity-dependent, spatially-dependent, and print-order biases**. *DNA Res* 2005, **12**(1):1-7.

61. Naderi A, Ahmed AA, Wang Y, Brenton JD, Caldas C: **Optimal amounts of fluorescent dye improve expression Microarray results in tumor specimens**. *Mol Biotechnol* 2005, **30**(2):151-154.

62. Bafico A, Varesco L, De Benedetti L, Caligo MA, Gismondi V, Sciallero S, Aste H, Ferrara GB, Bevilacqua G: **Genomic PCR-SSCP analysis of the metastasis associated NM23-H1 (NME1) gene: a study on colorectal cancer**. *Anticancer Res* 1993, **13**(6A):2149-2154.

63. Chandrasekharappa SC, Gross LA, King SE, Collins FS: **The human NME2 gene lies within 18kb of NME1 in chromosome 17**. *Genes Chromosomes Cancer* 1993, **6**(4):245-248.

64. Cropp CS, Lidereau R, Leone A, Liscia D, Cappa AP, Campbell G, Barker E, Le Doussal V, Steeg PS, Callahan R: **NME1 protein expression and loss of heterozygosity mutations in primary human breast tumors**. *J Natl Cancer Inst* 1994, **86**(15):1167-1169.

65. Lamb RF, Going JJ, Pickford I, Birnie GD: **Allelic imbalance at NME1 in microdissected primary and metastatic human colorectal carcinomas is frequent but not associated with metastasis to lymph nodes or liver**. *Cancer Res* 1996, **56**(4):916-920.

66. Leary JA, Kerr J, Chenevix-Trench G, Doris CP, Hurst T, Houghton CR, Friedlander ML: **Increased expression of the NME1 gene is associated with metastasis in epithelial ovarian cancer**. *Int J Cancer* 1995, **64**(3):189-195.

67. Miele ME, De La Rosa A, Lee JH, Hicks DJ, Dennis JU, Steeg PS, Welch DR: **Suppression of human melanoma metastasis following introduction of chromosome 6 is independent of NME1 (Nm23)**. *Clin Exp Metastasis* 1997, **15**(3):259-265.

68. Scholnick SB, Sun PC, Shaw ME, Haughey BH, el-Mofty SK: **Frequent loss of heterozygosity for Rb, TP53, and chromosome arm 3p, but not NME1 in squamous cell carcinomas of the supraglottic larynx**. *Cancer* 1994, **73**(10):2472-2480.

69. Toulas C, Mihura J, de Balincourt C, Marques B, Marek E, Soula G, Roche H, Favre G: **Potential prognostic value in human breast cancer of cytosolic Nme1 protein detection using an original hen specific antibody**. *Br J Cancer* 1996, **73**(5):630-635.

70. Boldrini L, Donati V, Dell'Omodarme M, Prati MC, Faviana P, Camacci T, Lucchi M, Mussi A, Santoro M, Basolo F *et al*: **Prognostic significance of osteopontin expression in early-stage non-small-cell lung cancer**. *Br J Cancer* 2005, **93**(4):453-457.

71. Donati V, Boldrini L, Dell'Omodarme M, Prati MC, Faviana P, Camacci T, Lucchi M, Mussi A, Santoro M, Basolo F *et al*: **Osteopontin expression and prognostic significance in non-small cell lung cancer**. *Clin Cancer Res* 2005, **11**(18):6459-6465.

72. Hu Z, Lin D, Yuan J, Xiao T, Zhang H, Sun W, Han N, Ma Y, Di X, Gao M *et al*: **Overexpression of osteopontin is associated with more aggressive phenotypes in human non-small cell lung cancer**. *Clin Cancer Res* 2005, **11**(13):4646-4652.

73. Hu Z, Xiao T, Lin DM, Guo SP, Zhang ZQ, Di XB, Cheng SJ, Gao YN: **[Over-expression of osteopontin in non-small cell lung cancers: its clinical significance]**. *Zhonghua Zhong Liu Za Zhi* 2007, **29**(8):591-595.

74. Le QT, Cao H, Koong A, Giaccia A: **Comment on: osteopontin as toxic marker**. *Radiother Oncol* 2006, **78**(2):230; author reply 230-231.

75. Schneider S, Yochim J, Brabender J, Uchida K, Danenberg KD, Metzger R, Schneider PM, Salonga D, Holscher AH, Danenberg PV: **Osteopontin but not osteonectin**

messenger RNA expression is a prognostic marker in curatively resected non-small cell lung cancer. *Clin Cancer Res* 2004, **10**(5):1588-1596.

Chapter 3:

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

2. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection**. *Proc Natl Acad Sci U S A* 2001, **98**(1):31-36.

3. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M *et al*: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proc Natl Acad Sci U S A* 2001, **98**(24):13790-13795.

4. Stearman RS, Dwyer-Nield L, Zerbe L, Blaine SA, Chan Z, Bunn PA, Jr., Johnson GL, Hirsch FR, Merrick DT, Franklin WA *et al*: **Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model**. *Am J Pathol* 2005, **167**(6):1763-1775.

5. Rosner B: **Fundamentals of Biostatistics**, 5th edn. Pacific Grove: Duxbury; 2000.

6. Alpaydin E: **Introduction to Machine Learning**. Cambridge: The MIT Press; 2004.

7. Manly BFJ: **Multivariate Statistical Methods**, 3rd edn. Washington D.C.: Chapman & Hall/CRC; 2005.

8. **Linear Discriminant Analysis, A Brief tutorial** [http://www.music.mcgill.ca/~ich/classes/mumt611/classifiers/lda_theory.pdf]

9. **Fisher Linear Discriminant Analysis** [http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf]

10. **Random Forest** [http://oz.berkeley.edu/users/breiman/randomforest2001.pdf]

11. **Looking Inside the Black Box** [http://stat-www.berkeley.edu/users/breiman/wald2002-2.pdf]

12. Diaz-Uriarte R, Alvarez de Andres S: **Gene selection and classification of Microarray data using random forest**. *BMC Bioinformatics* 2006, **7**:3.

13. Davis JG, M.: **The Relationship between Precision-Recall and ROC Curves**. *ICML Proceedings of the 23rd International Conference on Machine Learning* 2006:8.

14. Everson Richard M. JF: **Multi-class ROC analysis from a multi-objective optimisation perspective**. *Pattern Recognition Letters* 2006, **27**(8):22.

15. Irizarry RA: **affy**. In.: Bioconductor.

16. R DCT: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing.

17. Li C, Hung Wong W: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. *Genome biology* 2001, **2**(8):RESEARCH0032.

18. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.

19. Berrar DP, Downes CS, Dubitzky W: **Multiclass cancer classification using gene expression profiling and probabilistic neural networks**. *Pac Symp Biocomput* 2003:5-16.

20. Fridlyand SDaJ: **Introduction to Classification in Microarray Experiments**. In: *DNA Arrays Methods and Protocols.* Edited by Rampal JB, vol. 170. Totoja, NJ: Humana Press: 132-149.

21. G. Parmigiani ESG, R. A. Irizarry, S. L. Zeger: **The Analysis of Gene Expression Data**. New York: Springer; 2003.

22. Kostka D, Spang R: **Microarray based diagnosis profits from better documentation of gene expression signatures**. *PLoS Comput Biol* 2008, **4**(2):e22.

23. Li L, Weinberg CR, Darden TA, Pedersen LG: **Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method**. *Bioinformatics (Oxford, England)* 2001, **17**(12):1131-1142.

24. Ringner M, Peterson C: **Microarray-based cancer diagnosis with artificial neural networks**. *Biotechniques* 2003, **Suppl**:30-35.

25. Ringner M, Peterson C, Khan J: **Analyzing array data using supervised methods**. *Pharmacogenomics* 2002, **3**(3):403-415.

26. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S: **A comprehensive evaluation of multicategory classification methods for Microarray gene expression cancer diagnosis**. *Bioinformatics (Oxford, England)* 2005, **21**(5):631-643.

27. Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF: **GEMS: a system for automated cancer diagnosis and biomarker discovery from Microarray gene expression data**. *International journal of medical informatics* 2005, **74**(7-8):491-503.

28. Statnikov A, Wang L, Aliferis CF: **A comprehensive comparison of random forests and support vector machines for Microarray-based cancer classification**. *BMC Bioinformatics* 2008, **9**:319.

29. Szallasi Z: **Bioinformatics. Gene expression patterns and cancer**. *Nat Biotechnol* 1998, **16**(13):1292-1293.

30. Breiman L: **Bagging predictors**. *Machine Learning* 1996, **24**(2):18.

31. Yoav Freund RES: **A Short Introduction to Boosting**. *Journal of Japanese Society for Artificial Intelligence* 1999, **14**(5):10.

32. Wei C, Li J, Bumgarner RE: **Sample size for detecting differentially expressed genes in Microarray experiments**. *BMC Genomics* 2004, **5**(1):87.

33. **Online Handbook of Biological Statistics** [http://udel.edu/~mcdonald/statintro.html]

34. Deshmukh H: **Modeling the Physical Parameters Affecting the Measurements from Microarrays**. Fairfax: George Mason University; 2006.

35. Li L, Umbach DM, Terry P, Taylor JA: **Application of the GA/KNN method to SELDI proteomics data**. *Bioinformatics (Oxford, England)* 2004, **20**(10):1638-1640.

36. Ian H. Witten EF: **Data Mining: Practical Machine Learning Tools and Techniques**, 2nd edn: Morgan Kaufmann; 2005.

37. Ntzani EE, Ioannidis JP: **Predictive ability of DNA Microarrays for cancer outcomes and correlates: an empirical assessment**. *Lancet* 2003, **362**(9394):1439-1444.

38. Pocernich M: **verification**. In.: R Foundation for Statistical Computing: This package contains utilities for verification of discrete,continuous, probabilistic forecasts and forecast expressed as parametric distributions. .

39. Fodor AA, Tickle TL, Richardson C: **Towards the uniform distribution of null P values on Affymetrix Microarrays**. *Genome biology* 2007, **8**(5):R69.

40. Brody JP, Williams BA, Wold BJ, Quake SR: **Significance and statistical errors in the analysis of DNA Microarray data**. *Proc Natl Acad Sci U S A* 2002, **99**(20):12975-12978.

41. Edward Keedwell AN: **Intelligent Bioinformatics**
The application of artificial intelligence techniques to bioinformatics problems. Hoboken: John Wiley & Sons; 2005.

42. Ioannidis JP: **Microarrays and molecular research: noise discovery?** *Lancet* 2005, **365**(9458):454-455.

43. Shields R: **The emperor's new clothes revisited**. *Trends Genet* 2006, **22**(9):463.

Chapter 4:

1. Minna JD, Roth JA, Gazdar AF: **Focus on lung cancer**. *Cancer cell* 2002, **1**(1):49-52.

2. Travis WD, Travis LB, Devesa SS: **Lung cancer**. *Cancer* 1995, **75**(1 Suppl):191-202.

3. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M *et al*: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proc Natl Acad Sci U S A* 2001, **98**(24):13790-13795.

4. Stearman RS, Dwyer-Nield L, Zerbe L, Blaine SA, Chan Z, Bunn PA, Jr., Johnson GL, Hirsch FR, Merrick DT, Franklin WA *et al*: **Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model**. *Am J Pathol* 2005, **167**(6):1763-1775.

5. Shlens J: **A Tutorial on Principal Component Analysis**. In. La Jolle: Systems Neurobiology Laboratory, Salk Institue for Biological Studies/Institue for Nonlinear Science, UCSD; 2005: 11.

6. Fodor IK: **A Survey of Dimension Reduction Techniques**. In. Livermore, CA: US DOE Office of Scientific and Technical Information; 2002: 18.

7. Higgs BW, Weller J, Solka JL: **Spectral embedding finds meaningful (relevant) structure in image and Microarray data**. *BMC bioinformatics* 2006, **7**:74.

8. **Online Handbook of Biological Statistics** [http://udel.edu/~mcdonald/statintro.html]

9. Rosner B: **Fundamentals of Biostatistics**, 5th edn. Pacific Grove: Duxbury; 2000.

10. Alpaydin E: **Introduction to Machine Learning**. Cambridge: The MIT Press; 2004.

11. **Linear Discriminant Analysis, A Brief tutorial** [http://www.music.mcgill.ca/~ich/classes/mumt611/classifiers/lda_theory.pdf]

12. **Random Forest** [http://oz.berkeley.edu/users/breiman/randomforest2001.pdf]

13. **Fisher Linear Discriminant Analysis** [http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf]

14. Everson Richard M. JF: **Multi-class ROC analysis from a multi-objective optimisation perspective**. *Pattern Recognition Letters* 2006, **27**(8):22.

15. Davis JG, M.: **The Relationship between Precision-Recall and ROC Curves**. *ICML Proceedings of the 23rd International Conference on Machine Learning* 2006:8.

16. Hothorn T: **maxstat**. In.: R Foundation for Statistical Computing: Maximally selected rank statistics with several p-value approximations.

17. R DCT: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing.

18. **Survival Curves: The Basics** [http://cancerguide.org/scurve_basic.html]

19. Hu Z, Lin D, Yuan J, Xiao T, Zhang H, Sun W, Han N, Ma Y, Di X, Gao M *et al*: **Overexpression of osteopontin is associated with more aggressive phenotypes in human non-small cell lung cancer**. *Clin Cancer Res* 2005, **11**(13):4646-4652.

20. Donati V, Boldrini L, Dell'Omodarme M, Prati MC, Faviana P, Camacci T, Lucchi M, Mussi A, Santoro M, Basolo F *et al*: **Osteopontin expression and prognostic significance in non-small cell lung cancer**. *Clin Cancer Res* 2005, **11**(18):6459-6465.

21. Ntzani EE, Ioannidis JP: **Predictive ability of DNA Microarrays for cancer outcomes and correlates: an empirical assessment**. *Lancet* 2003, **362**(9394):1439-1444.

22. Le QT, Chen E, Salim A, Cao H, Kong CS, Whyte R, Donington J, Cannon W, Wakelee H, Tibshirani R *et al*: **An evaluation of tumor oxygenation and gene expression in patients with early stage non-small cell lung cancers**. *Clin Cancer Res* 2006, **12**(5):1507-1514.

23. Schneider S, Yochim J, Brabender J, Uchida K, Danenberg KD, Metzger R, Schneider PM, Salonga D, Holscher AH, Danenberg PV: **Osteopontin but not osteonectin messenger RNA expression is a prognostic marker in curatively resected non-small cell lung cancer**. *Clin Cancer Res* 2004, **10**(5):1588-1596.

24. Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF: **GEMS: a system for automated cancer diagnosis and biomarker discovery from Microarray gene expression data**. *International journal of medical informatics* 2005, **74**(7-8):491-503.

25. Statnikov A, Wang L, Aliferis CF: **A comprehensive comparison of random forests and support vector machines for Microarray-based cancer classification**. *BMC bioinformatics* 2008, **9**:319.

26. Alibes A, Canada A, Diaz-Uriarte R: **PaLS: filtering common literature, biological terms and pathway information**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W364-367.

27. Cao G, O'Brien CD, Zhou Z, Sanders SM, Greenbaum JN, Makrigiannakis A, DeLisser HM: **Involvement of human PECAM-1 in angiogenesis and in vitro endothelial cell migration**. *Am J Physiol Cell Physiol* 2002, **282**(5):C1181-1190.

28. Gergely F, Karlsson C, Still I, Cowell J, Kilmartin J, Raff JW: **The TACC domain identifies a family of centrosomal proteins that can interact with microtubules**. *Proc Natl Acad Sci U S A* 2000, **97**(26):14352-14357.

29. Gergely F: **Centrosomal TACCtics**. *Bioessays* 2002, **24**(10):915-925.

30. Delaval B, Ferrand A, Conte N, Larroque C, Hernandez-Verdun D, Prigent C, Birnbaum D: **Aurora B -TACC1 protein complex in cytokinesis**. *Oncogene* 2004, **23**(26):4516-4522.

31. Conte N, Charafe-Jauffret E, Delaval B, Adelaide J, Ginestier C, Geneix J, Isnardon D, Jacquemier J, Birnbaum D: **Carcinogenesis and translational controls: TACC1 is down-regulated in human cancers and associates with mRNA regulators**. *Oncogene* 2002, **21**(36):5619-5630.

Chapter 5:

1. Minna JD, Roth JA, Gazdar AF: **Focus on lung cancer**. *Cancer cell* 2002, **1**(1):49-52.

2. Travis WD, Travis LB, Devesa SS: **Lung cancer**. *Cancer* 1995, **75**(1 Suppl):191-202.

3. Alpaydin E: **Introduction to Machine Learning**. Cambridge: The MIT Press; 2004.
4. Edward Keedwell AN: **Intelligent Bioinformatics**
The application of artificial intelligence techniques to bioinformatics problems. Hoboken: John Wiley & Sons; 2005.

5. Fridlyand SDaJ: **Introduction to Classification in Microarray Experiments**. In: *DNA Arrays Methods and Protocols.* Edited by Rampal JB, vol. 170. Totoja, NJ: Humana Press: 132-149.

6. Manly BFJ: **Multivariate Statistical Methods**, 3rd edn. Washington D.C.: Chapman & Hall/CRC; 2005.

7. **Linear Discriminant Analysis, A Brief tutorial**
[http://www.music.mcgill.ca/~ich/classes/mumt611/classifiers/lda_theory.pdf]

8. **Fisher Linear Discriminant Analysis**
[http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf]

9. R DCT: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing.

10. Ian H. Witten EF: **Data Mining: Practical Machine Learning Tools and Techniques**, 2nd edn: Morgan Kaufmann; 2005.

11. Breiman L: **Bagging predictors**. *Machine Learning* 1996, **24**(2):18.

12. Yoav Freund RES: **A Short Introduction to Boosting**. *Journal of Japanese Society for Artificial Intelligence* 1999, **14**(5):10.

13. Davis JG, M.: **The Relationship between Precision-Recall and ROC Curves**. *ICML Proceedings of the 23rd International Conference on Machine Learning* 2006:8.

14. Everson Richard M. JF: **Multi-class ROC analysis from a multi-objective optimisation perspective**. *Pattern Recognition Letters* 2006, **27**(8):22.

15. Delaval B, Ferrand A, Conte N, Larroque C, Hernandez-Verdun D, Prigent C, Birnbaum D: **Aurora B -TACC1 protein complex in cytokinesis**. *Oncogene* 2004, **23**(26):4516-4522.

16. Conte N, Delaval B, Ginestier C, Ferrand A, Isnardon D, Larroque C, Prigent C, Seraphin B, Jacquemier J, Birnbaum D: **TACC1-chTOG-Aurora A protein complex in breast cancer**. *Oncogene* 2003, **22**(50):8102-8116.

17. Conte N, Charafe-Jauffret E, Delaval B, Adelaide J, Ginestier C, Geneix J, Isnardon D, Jacquemier J, Birnbaum D: **Carcinogenesis and translational controls: TACC1 is down-regulated in human cancers and associates with mRNA regulators**. *Oncogene* 2002, **21**(36):5619-5630.

18. Gergely F, Karlsson C, Still I, Cowell J, Kilmartin J, Raff JW: **The TACC domain identifies a family of centrosomal proteins that can interact with microtubules**. *Proc Natl Acad Sci U S A* 2000, **97**(26):14352-14357.

19. Gergely F: **Centrosomal TACCtics**. *Bioessays* 2002, **24**(10):915-925.

20. Forbes A, Wadehra M, Mareninov S, Morales S, Shimazaki K, Gordon LK, Braun J: **The tetraspan protein EMP2 regulates expression of caveolin-1**. *J Biol Chem* 2007, **282**(36):26542-26551.

21. Ho CC, Huang PH, Huang HY, Chen YH, Yang PC, Hsu SM: **Up-regulated caveolin-1 accentuates the metastasis capability of lung adenocarcinoma by inducing filopodia formation**. *Am J Pathol* 2002, **161**(5):1647-1656.

22. Govers R, van der Sluijs P, van Donselaar E, Slot JW, Rabelink TJ: **Endothelial nitric oxide synthase and its negative regulator caveolin-1 localize to distinct perinuclear organelles**. *J Histochem Cytochem* 2002, **50**(6):779-788.

23. Mundy DI, Machleidt T, Ying YS, Anderson RG, Bloom GS: **Dual control of caveolar membrane traffic by microtubules and the actin cytoskeleton**. *J Cell Sci* 2002, **115**(Pt 22):4327-4339.

24. Thomas CM, Smart EJ: **Caveolae structure and function**. *J Cell Mol Med* 2008, **12**(3):796-809.

25. Goligorsky MS, Li H, Brodsky S, Chen J: **Relationships between caveolae and eNOS: everything in proximity and the proximity of everything**. *Am J Physiol Renal Physiol* 2002, **283**(1):F1-10.

26. Yano S, Yano N: **Regulation of catalase enzyme activity by cell signaling molecules**. *Mol Cell Biochem* 2002, **240**(1-2):119-130.

27. Bai J, Cederbaum AI: **Catalase protects HepG2 cells from apoptosis induced by DNA-damaging agents by accelerating the degradation of p53**. *J Biol Chem* 2003, **278**(7):4660-4667.

28. Kondo T, Kitano T, Iwai K, Watanabe M, Taguchi Y, Yabu T, Umehara H, Domae N, Uchiyama T, Okazaki T: **Control of ceramide-induced apoptosis by IGF-1: involvement of PI-3 kinase, caspase-3 and catalase**. *Cell Death Differ* 2002, **9**(6):682-692.

29. Oshiro MM, Futscher BW, Lisberg A, Wozniak RJ, Klimecki WT, Domann FE, Cress AE: **Epigenetic regulation of the cell type-specific gene 14-3-3sigma**. *Neoplasia* 2005, **7**(9):799-808.

30. Yang HY, Wen YY, Chen CH, Lozano G, Lee MH: **14-3-3 sigma positively regulates p53 and suppresses tumor growth**. *Mol Cell Biol* 2003, **23**(20):7096-7107.

31. Zhang Y, Karas M, Zhao H, Yakar S, LeRoith D: **14-3-3sigma mediation of cell cycle progression is p53-independent in response to insulin-like growth factor-I receptor activation**. *J Biol Chem* 2004, **279**(33):34353-34360.

32. Dubois T, Paleotti O, Mironov AA, Fraisier V, Stradal TE, De Matteis MA, Franco M, Chavrier P: **Golgi-localized GAP for Cdc42 functions downstream of ARF1 to control Arp2/3 complex and F-actin dynamics**. *Nat Cell Biol* 2005, **7**(4):353-364.

33. Kaneda A, Kaminishi M, Sugimura T, Ushijima T: **Decreased expression of the seven ARP2/3 complex genes in human gastric cancers**. *Cancer Lett* 2004, **212**(2):203-210.

34. Taira K, Umikawa M, Takei K, Myagmar BE, Shinzato M, Machida N, Uezato H, Nonaka S, Kariya K: **The Traf2- and Nck-interacting kinase as a putative effector of Rap2 to regulate actin cytoskeleton**. *J Biol Chem* 2004, **279**(47):49488-49496.

35. Heale JT, Ball AR, Jr., Schmiesing JA, Kim JS, Kong X, Zhou S, Hudson DF, Earnshaw WC, Yokomori K: **Condensin I interacts with the PARP-1-XRCC1 complex and functions in DNA single-strand break repair**. *Mol Cell* 2006, **21**(6):837-848.

36. Maeda Y, Hunter TC, Loudy DE, Dave V, Schreiber V, Whitsett JA: **PARP-2 interacts with TTF-1 and regulates expression of surfactant protein-B**. *J Biol Chem* 2006, **281**(14):9600-9606.

37. Brasch F, Johnen G, Winn-Brasch A, Guttentag SH, Schmiedl A, Kapp N, Suzuki Y, Muller KM, Richter J, Hawgood S *et al*: **Surfactant protein B in type II pneumocytes and intra-alveolar surfactant forms of human lungs**. *Am J Respir Cell Mol Biol* 2004, **30**(4):449-458.

38. Das A, Boggaram V: **Proteasome dysfunction inhibits surfactant protein gene expression in lung epithelial cells: mechanism of inhibition of SP-B gene expression**. *Am J Physiol Lung Cell Mol Physiol* 2007, **292**(1):L74-84.

39. Sparkman L, Chandru H, Boggaram V: **Ceramide decreases surfactant protein B gene expression via downregulation of TTF-1 DNA binding activity**. *Am J Physiol Lung Cell Mol Physiol* 2006, **290**(2):L351-358.

40. Wright JR: **Immunoregulatory functions of surfactant proteins**. *Nat Rev Immunol* 2005, **5**(1):58-68.

41. Yang L, Yan D, Yan C, Du H: **Peroxisome proliferator-activated receptor gamma and ligands inhibit surfactant protein B gene expression in the lung**. *J Biol Chem* 2003, **278**(38):36841-36847.

42. Kargi A, Gurel D, Tuna B: **The diagnostic value of TTF-1, CK 5/6, and p63 immunostaining in classification of lung carcinomas**. *Appl Immunohistochem Mol Morphol* 2007, **15**(4):415-420.

43. Stefansson IM, Salvesen HB, Akslen LA: **Loss of p63 and cytokeratin 5/6 expression is associated with more aggressive tumors in endometrial carcinoma patients**. *Int J Cancer* 2006, **118**(5):1227-1233.

44. Swarbrick A, Akerfeldt MC, Lee CS, Sergio CM, Caldon CE, Hunter LJ, Sutherland RL, Musgrove EA: **Regulation of cyclin expression and cell cycle progression in breast epithelial cells by the helix-loop-helix protein Id1**. *Oncogene* 2005, **24**(3):381-389.

45. Zhang X, Ling MT, Wang Q, Lau CK, Leung SC, Lee TK, Cheung AL, Wong YC, Wang X: **Identification of a novel inhibitor of differentiation-1 (ID-1) binding partner, caveolin-1, and its role in epithelial-mesenchymal transition and resistance to apoptosis in prostate cancer cells**. *J Biol Chem* 2007, **282**(46):33284-33294.

46. Korchynskyi O, ten Dijke P: **Identification and functional characterization of distinct critically important bone morphogenetic protein-specific response elements in the Id1 promoter**. *J Biol Chem* 2002, **277**(7):4883-4891.

47. Pruneri G, Pignataro L, Valentini S, Fabris S, Maisonneuve P, Carboni N, Pece S, Capra M, Del Curto B, Neri A *et al*: **Cyclin D3 immunoreactivity is an independent predictor of survival in laryngeal squamous cell carcinoma**. *Clin Cancer Res* 2005, **11**(1):242-248.

48. Zhao Y, Hamza MS, Leong HS, Lim CB, Pan YF, Cheung E, Soo KC, Iyer NG: **Kruppel-like factor 5 modulates p53-independent apoptosis through Pim1 survival kinase in cancer cells**. *Oncogene* 2008, **27**(1):1-8.

49. Nandan MO, Chanchevalap S, Dalton WB, Yang VW: **Kruppel-like factor 5 promotes mitosis by activating the cyclin B1/Cdc2 complex during oncogenic Ras-mediated transformation**. *FEBS Lett* 2005, **579**(21):4757-4762.

50. Li HH, Cai X, Shouse GP, Piluso LG, Liu X: **A specific PP2A regulatory subunit, B56gamma, mediates DNA damage-induced dephosphorylation of p53 at Thr55**. *Embo J* 2007, **26**(2):402-411.

51. Shouse GP, Cai X, Liu X: **Serine 15 phosphorylation of p53 directs its interaction with B56gamma and the tumor suppressor activity of B56gamma-specific protein phosphatase 2A**. *Mol Cell Biol* 2008, **28**(1):448-456.

CURRICULUM VITAE

Kevin Thompson received his Bachelor of Science in Biology from the University of Wisconsin-Stevens Point in 1995. He worked in the research field for 5 years; most of that time was with Dr. Micheal Mosesson's Laboratory and completed his Master's of Science in MIS in 2002, from UW-Milwaukee.