VARIATIONS IN MICROARRAY BASED GENE EXPRESSION PROFILING: IDENTIFYING SOURCES AND IMPROVING RESULTS

by

Changqing Ma

Bachelor of Medicine, Beijing Medical University, 1998

MSIS, University of Pittsburgh, 2001

Submitted to the Graduate Faculty of

The School of Medicine, Department of Pathology in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

FACULTY OF MEDICINE

This dissertation was presented

by

Changqing Ma

It was defended on

July 8th 2005

and approved by

John Gilbertson, MD

George Michalopoulos, MD PhD

Xiao-ming Yin, MD PhD

Vanathi Gopalakrishnan, PhD

Michael Becich, MD PhD
Dissertation Director

VARIATIONS IN MICROARRAY BASED GENE EXPRESSION PROFILING: IDENTIFYING SOURCES AND IMPROVING RESULTS

Changqing Ma PhD

University of Pittsburgh, 2005

ABSTRACT

Two major issues hinder the application of microarray based gene expression profiling in clinical laboratories as a diagnostic or prognostic tool. The first issue is the sheer volume and high-dimensionality of gene expression data from microarray experiments, which require advanced algorithms to extract meaningful gene expression patterns that correlate with biological impact. The second issue is the substantial amount of variation in microarray gene expression data, which impairs the performance of analysis method and makes sharing or integrating microarray data very difficult. Variations can be introduced by all possible sources including the DNA microarray technology itself and the experimental procedures. Many of these variations have not been characterized, measured, or linked to the sources.

In the first part of this dissertation, a decision tree learning method was demonstrated to perform as well as more popularly accepted classification methods in partitioning cancer samples with microarray data. More importantly, results demonstrate that variation introduced into microarray data by tissue sampling and tissue handling compromised the performance of classification methods.

In the second part of this dissertation, variations introduced by the T7 based *in vitro* transcription labeling methods were investigated in detail. Results demonstrated that individual amplification methods significantly biased gene expression data even though the methods compared in this study were all derivatives of the T7 RNA polymerase based *in vitro* transcription labeling approach. Variations observed can be partially explained by the number of biotinylated nucleotides used for labeling and the incubation time of the *in vitro* transcription experiments. These variations can generate discordant gene expression results even using the same RNA samples and cannot be corrected by post experiment analysis including advanced normalization techniques.

Studies in this dissertation stress the concept that experimental and analytical methods must work together. This dissertation also emphasizes the importance of standardizing the DNA microarray technology and experimental procedures in order to optimize gene expression analysis and create quality standards compatible with the clinical application of this technology. These findings should be taken into account especially when comparing data from different platforms, and in standardizing protocols for clinical applications in pathology.

PREFACE

The successful application of DNA microarray based gene expression profiling in translational research makes it a potential tool for use in clinical laboratories for diagnostic, prognostic, and therapeutic applications in the era of molecular medicine. However, the DNA microarray must overcome two significant hurdles before it can be used use in clinical environment. First, the massive volume of the gene expression data from DNA microarray based experiments requires a set of reliable algorithms that can discover patterns of gene expression robustly with high sensitivity and specificity. Second, published results from many studies have demonstrated the existence of significant variations in the data introduced by both the technology itself and the experimental procedures utilized to produce gene expression data. These variations can significantly impair the utility of microarray data in identifying biologically meaningful gene expression patterns.

This dissertation represents work over five years, a time span in which research in microarray data has migrated from statistical analysis of large high-dimensionality data set to the understanding and control of experimental data variation between platforms, laboratories, and experiments. For these reasons, this dissertation has two main objectives.

The first objective is to test the usefulness of a decision-tree learning algorithm for classification using large high-dimensionality gene expression data sets from DNA microarray experiments. Results and implication of this work are discussed in Chapter II. The second objective is to study the experimental variation introduced by a particular RNA labeling method and discovery its source. Results from this study are described in Chapter IV of this dissertation. In addition, Chapter III reports results from an attempt to integrate gene expression data generated from different types of arrays at different institutions. This work contributes to our overall understanding of other sources of variation in the DNA microarray technology and microarray based gene expression profiling experiments.

Two other studies are included as appendixes. Appendix A is a study describing the gene expression patterns in different types of prostate tissue specimens, which was done in collaboration with others in the Becich laboratories and directly related to the theme of this dissertation. Appendix B is a manuscript submitted to the Intelligence System for Molecular Biology (ISMB) 2002 Conference as a student paper. It was presented at the conference as a poster. It reports the results of using decision-tree learning to detect the global gene expression changes in rat brain after cocaine treatment. This study was collaboration with Drs David G. Peter and Robert E. Ferrell from Department of Human Genetics and Dr. Vanathi Gopalakrishnan from the Center for Biomedical Informatics (CBMI). It also is the initial phase of applying the decision-tree learning algorithm on microarray data which was tested in Chapter II of this dissertation.

Chapter I summarizes the background knowledge related to the results presented in Chapter II, III, and IV. Section 1.1 provides an overview of the DNA microarray technology and discusses in detail the strategies of probe design and selection used for producing the Affymetrix GeneChip® arrays. It will help better understand the platform-dependent strategy proposed in Chapter III (Section 3.3.2.2) for microarray data integration. It will also help to understand how gene expression intensity values from the Affymetrix GeneChip$^{®}$ arrays are calculated by various algorithms in Section 1.3.2. Section 1.2 reviews the experimental procedures of DNA microarray based gene expression profiling experiments. RNA labeling methods are discussed in detail because they are important to understand the results presented in Chapter IV. Section 1.3-1.5 review the analysis of gene expression data from DNA microarray experiments. This includes both low-level and high-level analysis, each of which involves different principles. Both low-level and high-level data analysis are applied for every study conducted in this dissertation.

Finally, Section 1.5 provides a thorough discussion of experimental variation in gene expression data in DNA microarray based experiments. Awareness of the extreme importance of this topic has increased recently and many manuscripts have reported on sources of data variation in microarray experiments. This section provides a systematic review of the current understanding of the problem and the possible solutions proposed. In particular, variations related to RNA labeling methods, tissue sampling and handling are discussed in detail as they provide direct background for results in Chapter II, III, and IV respectively.

DEDICATION

This dissertation is dedicated to my loving father (Dabang Ma), mother (Zongqing Yu), and sister (Wenyu Ma) for their enduring love and support throughout my life.

In memory of my beloved uncle Zongbai Yu (1935~2004).

ACKNOWLEDGEMENTS

I would like to thank my major advisor Dr. Michael Becich for his continuous support and inspiration over the past 6 years of my Ph.D. study. Dr. Becich is not only a great scientist but also a very caring advisor. Through these years, he has helped me in so many ways to get what I need to accomplish this thesis work. He also provides me guidance on my future career. I feel so grateful to have the opportunity to work in his group and learn from him. What he taught me will surely inspire and benefit me throughout my life.

I would like to thank my dissertation committee members, Drs Michael Becich, John Gilbertson, George Michalopoulos, Xiao-ming Yin, and Vanathi Gopalakrishnan, for their guidance, and contribution culminating in this work.

I would also like to offer my gratitude to my committee chair Dr. John Gilbertson for all the guidance and help. His encouragement and understanding helped me to get through many difficulties and frustrations. Without his support and patience, this thesis could not have been finished.

I would especially like to thank Dr. Federico Monzon-Bordonaba in Department of Pathology for offering me the opportunity to conduct microarray experiments in the Clinical Genomics Facility at University of Pittsburgh Cancer Institute. Results from those experiments were summarized in Chapter IV. At the time I was working in the lab, Dr. Federico Monzon-Bordonaba provide me many guidance and inspirations. These hands-on experiences helped me to understand the DNA microarray technology better.

I would also like to thank the past and current members of the Clinical Genomics Facility for their numerous help: Wenjing Liang, Maureen Lyons-Weiler, Christin Sciulli.

I would also like to thank Drs William LaFramboise and James Lyons-Weiler for their thoughtful suggestion to this work.

Last (and surely not least), I would like to offer my gratitude to Dr. Uma Chandran and Songhui Li at Center for Pathology Informatics, who have become good friends through these years. They have offered me much support and suggestions whenever I needed assistance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. CHAPTER I Introduction

The classical dogma of molecular biology[1] states that the genetic information stored in DNA flows to RNA and then from RNA to protein by means of transcription and translation respectively. Gene expression is the process by which genetic information at the DNA level is converted into functional proteins. The transcription of messenger RNA from DNA molecules is an important regulatory point in this process and may signal cascades of many other events. The complexity and abundance of the messenger RNA population in a cell or organ reflect the cellular events in response to the environmental changes. Therefore, the study of patterns of gene expression at the messenger RNA level under different physiological conditions will provide evidence for understanding of many biological systems and gene function.

In general, the study of gene expression involves the comparison of mRNA populations between two samples taken under different conditions such as diseased versus healthy or treated versus untreated. In the past two decades, gene expression analysis has evolved from studying only one differentially expressed gene at a time to a detailed survey of the whole transcriptome. This rapid progress is in large part driven by the development and application of DNA microarray techniques.

1.1.DNA microarray overview

1.1.1. Overview

What is a DNA microarray?

A DNA microarray is a small analytical device that holds hundreds or thousands of DNA molecules for the simultaneous examination of fluorescently labeled samples (cRNA, cDNA, mRNA or total RNA) prepared from the messenger RNA population taken from cell cultures or tissue samples[2-6]. The DNA molecules on array surfaces are called 'probes'; the fluorescently labeled samples are referred to as 'targets'. Probes are manufactured in a high-density fashion on a small area (1.28cm × 1.28 cm for Affymetrix GeneChip® array and 3 inch × 1 inch for typical cDNA microarray) of a flat and solid glass surface. Probes are ordered in grid and each probe on an array has a unique 'address' (or X-Y coordinate) on a two-dimensional surface. This ensures the correct identification of each probe by computer-controlled robots used in microarray production and by software tools for data analysis purpose. Each probe is also highly specific to identify only one sequence in the transcriptome. DNA microarray technology allows sifting through and analyzing genomic information with exceptional speed and precision compared to other existing methods such as Northern blot analysis, differential display or serial analysis of gene expression(SAGE)[2-4].

History about the development of DNA microarray technology

Early forms of DNA arrays were initially generated by spotting the bacterial colonies on filter membranes for gene identification and DNA sequencing studies. Subsequent improvements in laboratory automation enabled the creation of high-density filter arrays[7, 8]. Beginning in the 1980s, many studies used the high-density filter arrays for sequencing, analyzing different gene expression and identifying new genes[7-14]. These arrays are, in general, entitled as 'macroarrays' because of the 'gigantic' size (usually 8cm to 22cm in diameter) of the nylon filter membranes. Along with the advent of automatic high speed spotting techniques for array manufacture were the introductions of glass surface for hybridization by several research groups in the late 1980s and early 1990s[15-19]. These experiments established the feasibility of hybridizing on glass surface and further founded the array fabrication techniques currently in use today. In 1995, Schena et al.[4] from Stanford University Medical Center was the first to report the usage of cDNA microarray for quantitative monitoring of gene expression patterns. Then in 1996, Lockhart et al.[2] at Affymetrix Inc. (Santa Clara, CA) reported, for the first time, expression monitoring experiments by hybridization to high-density oligonucleotide arrays. Since then, cDNA microarrays and oligonucleotide arrays have become the major microarray platforms used for gene expression profiling. Based on a database analysis of 2,000 microarray citations from 1995 to 2002[3], 65% of all the publications based on DNA microarrays experiments used cDNA microarrays and 26% of them applied oligonucleotide microarrays in their studies. In addition, 81.5% of microarray publications are for gene expression analysis.

Type of DNA microarrays

One major difference between cDNA and oligonucleotide microarrays is the type of DNA probe utilized. Probes on cDNA microarrays are typically 500~2500bp double-strand cDNAs produced by PCR amplification of cDNA libraries[3-5, 20]. In contrast, probes on oligonucleotide microarrays are single-strand 20~90 nucleotide molecules synthesized *in vitro*. In addition, microarrays can also be categorized based on array fabrication methods. For spotted arrays, DNA probes are mechanically deposited on the array surface[3-5, 20-26]. For *in situ* arrays, probes are synthesized in silico[2, 3, 6, 21, 22, 27-29], such as the GeneChip® arrays manufactured by Affymetrix Inc.. (Santa Clara, CA) DNA microarrays may also be categorized as one- or two-channel format where the difference is at the number of dyes used to label targets. For a DNA microarray using two dyes, two channels are needed in the image acquisition devices to gather signal from each dye.

To date many species have been studied by microarray based experiments. Among all publications on microarrays from 1995 to 2002, more than half of microarray publications are from studies on human. Other popular species are mouse, yeast, and rat with decedent order of publications. *E. coli*, Arabidopsis, fruit fly, and *C. elegans* together account for 11.2% of all these publications. Other microarray-available species are zebra fish, *B. subtilis*, bovine, maize, P. aeruginosa, Plasmodium, Procine, *S. aureus*, soybean, sugar cane, wheat, Xenopus laevis etc. In theory, any organism/species can be studied by microarray; this is an attractive feature of microarray technology[3].

Applications

DNA microarray based gene expression profiling has been widely applied to all kinds of research involving gene expression profiling. The use of this technology for non-human species is focused on gene function, development and expression survey of the whole genome[29-31]. The most common use of microarrays is gene expression profiling of human specimens and, among those, the most popular application is gene expression profiling of human cancers. Many studies have presented results on discovery of gene functions, drug targets, pathway dissection, etc… More importantly, DNA microarray base gene expression profiling has been successfully applied in clinical research on the classification of clinical samples, discovery of subclasses of disease, and prediction of disease outcome and patient survival (also see Section 1.4.3 and Section 104.4)[32-40].

Other types of microarray

In addition to DNA microarrays, other types of microarrays have also been developed more recently for high-throughput profiling other types of molecules or biological systems. Among these, tissue microarray is worthy to be summarized.

Tissue microarray (TMA) technology[41] is used for high-throughput in-situ tissue analysis including immunohistochemistry (IHC), fluorescence *in situ* hybridization (FISH), mRNA *in situ* hybridization (mRNA ISH). A tissue microarray contains up to thousands of different tissue samples aligned in grid on a microscope glass slide. Each tissue sample on a TMA is usually 0.1 to 0.6mm in diameter; adjacent tissue samples are 0.1mm apart[41, 42]. Tissue samples on a TMA are from representative regions of original tissue samples fixed by formalin and embedded in paraffin blocks or fixed by cold ethanol and embedded in Tissue Tek® O.C.T.™ compound for preserving intact DNA and RNA[43]. As pointed out by Simon et al.[42], the process of making a TMA involves handling many tissue specimens as well as associated data. These include the identification of relevant samples from a tissue bank, the collection of glass slides for selected cases, the selection of morphologically representative area on slides, the collection of paraffin blocks of the selected cases and the storage of histological and clinical information of selected cases. TMAs can be constructed manually or using semi-automated devices with two needles. One needle punches a hole, 0.1 to 0.6mm in diameter, on the 'recipient' paraffin block at a specific coordinate. The other needle with inner diameter of 0.1 to 0.6mm retrieves tissue samples from the selected region on the 'donor' paraffin embedded tissue specimen and then precisely arrays the core biopsy in the pre-made hole. A 'donor' tissue specimen can provide many tissue biopsies with negligible damage. A recipient paraffin block can hold hundreds, up to thousands of core tissues. Glass slides containing very large number of tissue samples are made from consecutive sections cut from the 'recipient' block, and are hence called tissue microarrays or TMAs.

There are many advantages of using TMAs. The key advantage is that molecular markers can be examined in the context of tissue morphology. TMA technology provides a platform to detect DNA, RNA, or protein targets on a large number of different tissue types using uniform methodologies and interpretation criteria. As the original tissue specimens remain intact, TMA allows intensive studying small tissue specimens as well as preserve precious tissues for future investigations. TMAs also allow possible automated analysis of arrayed tissue samples as each of them has relatively precise position on a TMA[42, 44].

Although TMAs can be used for any types of *in situ* tissue analysis, TMAs have been utilized mainly in cancer research. Applications of TMAs have been very well reviewed with great details by Simon et al.[42, 44] and summarized as follows. TMAs can be constructed to include tissue samples from multiple tumor types. This type of TMAs has been utilized to evaluate the prevalence of markers in different types of tumors. TMAs can contain different stages of a particular type of tumor for studying tumor progression and detect associations between tumor phenotype and genotype. There are also TMAs containing tumor tissue samples with clinical follow-up data for studying cancer prognosis and reveal association between genetic alteration and clinical outcome[45].

1.1.2. DNA microarray fabrication

1.1.2.1. Overview

DNA microarrays are fabricated by various techniques including techniques for *in situ* synthesis of oligo-nucleotide on an array surface[2, 6, 15, 27, 28, 46-48] and techniques for spotting pre-synthesized probes onto array surfaces[3, 4, 23, 25]. DNA microarrays created by *in situ* methods, such as GeneChip® arrays from Affymetrix Inc., are only available commercially. Spotted DNA microarrays are commercially available but also can be manufactured at microarray facilities in academic centers such as the Brown's lab at Stanford University (http://brownlab.stanford.edu/).

Microarray fabrication begins with the selection of a panel of appropriate probes that will be attached on the array surface. Two types of probes are currently in use for DNA microarray based gene expression profiling: cDNA probes and oligonucleotide probes. There are two approaches to prepare probes for DNA microarray manufacture: (1) cDNA probes and oligonucleotide probes can be prepared beforehand and then delivered onto the array surface; (2) Oligonucleotide probes can be synthesized *in situ* during array manufacture (refers to the *in situ* synthesis mentioned above). Accordingly, tasks related to probe preparation also vary. The major tasks in probe preparations include design, selection, and annotation of probes. Figure 1.1 outlines the major steps/tasks for probe preparation based on probe type.

Oligonucleotide probes may be synthesized *in situ* or fabricated prior array construction (Figure 1.1a). Either way, sequence information is required to design and prepare oligonucleotide probes. Since the recent completion of sequencing the whole genomes of human and other species[30, 31, 49], the availability of sequence information is no longer a factor that limits the production of oligonucleotide arrays. Probes are selected from established sequence databases, such as UniGene[50] and TIGR Gene Indices[51]. After a region on the sequence of a transcript is selected as a potential probe to represent the transcript on a microarray, this region needs to be validated by comparing it with all sequences, if available, in the transcriptome of the studied organism to minimize cross-hybridization using sequence similarity algorithms such as global BLAST. If the region is not unique enough to identify single transcript in the transcriptome, it is dropped and another region in the sequence of the same transcript may be selected and evaluated using the same strategy until a unique region is identified. Once probes are selected, precursors are prepared for microarray construction *in situ*, or probes are prepared using classical phosphoramidite chemistry[3, 52, 53] and purified before delivering to the array surface(Figure 1.1a).

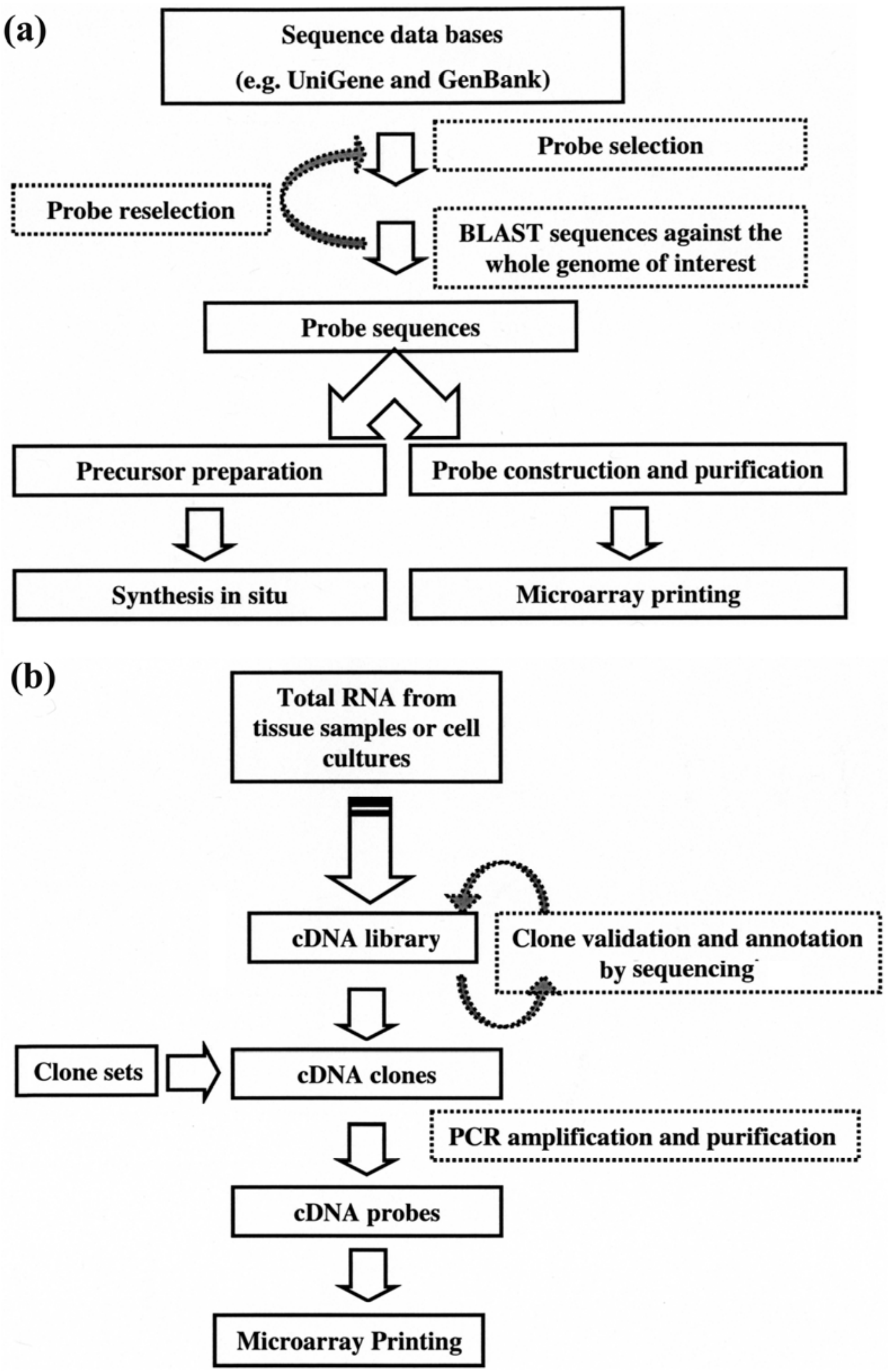

(a)
Sequence data bases
(e.g. UniGene and GenBank)
Probe selection
Probe reselection
BLAST sequences against the whole genome of interest
Probe sequences
Precursor preparation
Probe construction and purification
Synthesis in situ
Microarray printing
(b)
Total RNA from tissue samples or cell cultures
cDNA library
Clone validation and annotation by sequencing
Clone sets
cDNA clones
PCR amplification and purification
cDNA probes
Microarray Printing

Figure 1.1 Schematic of probe perpetration workflow for Oligonucleotide microarrays (a) and cDNA microarrays (b).

cDNA probes are obtained by purifying PCR products of cDNA clones (Figure 1.1 b). In contrast to oligonucleotide probes, the availability of sequence information is not a prerequisite for cDNA probe preparation. cDNA clones usually can be obtained as pre-constructed, validated, and annotated clone sets from academic and commercial resources. Under certain circumstances, such as the lack of clone sets for the organism of interest, a cDNA library may need to be prepared from a specific cell type or tissue[54]. Before cDNA clones can be used for cDNA probe amplification by PCR, each individual clone in the library needs to be sequenced to verify the sequence itself and avoid possible redundancy. cDNA clones are selected to represent as many unique transcripts as possible to produce arrays with low redundancy of transcript representation to survey the broadest possible set of genes.

1.1.2.2. Affymetrix strategies for probe design and selection

Affymetrix has unique proprietary ways to design and selection probes for gene expression analysis. (Please refer to Lipshutz et al.[6] for a graphical illustration.) This design strategy determines that hybridization results using Affymetrix GeneChip® arrays need to be analyzed by specially designed algorithms (discussed in Section 1.3.2). On the other hand, this strategy provide means by which platform-dependent integration approach can be developed as presented Chapter III for the integration of microarray gene expression data from different generations of the Affymetrix GeneChip® arrays.

When selecting probes from sequence databases, a set of heuristic rules and/or model-based approach is applied to identify candidate sequences as probes. Significantly, redundancy is applied to every single transcript surveyed by the array[2, 6, 29, 55, 56]. Multiple probes (11~20) are used for each transcript. This redundant set of probes is called a “probe set”. These approaches lead to the current set of GeneChip® arrays for gene expression analysis[6].

Probe redundancy

A set of probes, named a "probe set," is used to represent one transcript. A probe set comprises 11 to 20 independent pairs of probes/oligonucleotides. Each probe matches to a sequence region of the transcript with minimum, if not, overlapping to other probe sets. In the Affymetrix design, each probe is actually implemented as a probe pair. A probe pair consists of a perfect match probe (PM) and a miss match probe (MM) and both PM and MM probes are 25-mer in length. The PM probe is 100% complementary to target sequence; the MM probe, on the other hand, has a single mismatch at the 13th base. Therefore, the only difference between PM and MM sequence is in the central position of the sequence[2, 6, 29, 55, 56].

To summarize, in the Affymetrix GeneChip® arrays, redundancy is present at two levels: (1) the use of multiple oligonucleotides with different sequences (a probe set) to hybridize on different regions of a single transcript, (2) the use of MM probes in addition to the PM probe. These redundancies may offer advantages in the context of gene expression detection. When multiple oligonucleotides are used for the same transcript, the signal is derived by taking an average of the set of individual hybridization signal from each oligonucleotide. This may provide a better signal, decrease the signal-to-noise ratio, and increase accuracy for RNA quantitation. MM probes can be used as a control for cross-hybridization and non-specific hybridization. When calculating intensity value for a target, intensity values from MM probes in its probe set will be subtracted. At low concentration of a target, hybridization to the PM and MM probe pair will help to distinguish true signal with background/noise, i.e. whether the signal is generated by hybridizing the right target on its probe set or from non-specific hybridization of any type[6].

Probe selection rules

Each probe set representing a transcript is selected from 3' end sequence region of its intended target. For example, regardless the length of the transcripts, the maximum distance of the 5'-most probe pair to the 3' end of the transcript on the HG-U95Av2 array and HG-U133 set is about 600 base pair. This 3'-biased probe selection strategy ensures maximum number of targets are available for detection even when mRNA have been partially degraded because mRNA degradation usually starts from the 5' end. Uniqueness of probes is assessed by comparing the candidate probe sequence first to all probes on the array and then to the full-length sequences of all transcripts in the surveyed species, if available. Probes were rejected if there were 22 or more base positions matched. Further selection is based on the performance of probes in hybridization experiments. A set of heuristic rules have been developed for probe set selection based on probe behavior as a function of certain sequence features[2]. Neural network algorithms are then used to assess probe characteristics. At the last step, probes were rejected if more than 60 synthesis steps are needed to help minimize synthesis time and cost[2, 29].

1.2.DNA microarray based gene expression profiling

1.2.1. Experimental procedure

The experimental procedure of a gene expression profiling study using DNA microarrays is summarized as follows (Figure 1.2 a). The first step in such a study is to identify the sources of RNA. This is usually done while designing the study and the source is determined by the study objectives. Most common sources of RNA are either cell cultures or tissue specimens from experiment animals or patients. Experiments start from collecting RNA samples by extracting them from the RNA sources. Targets are then prepared by labeling RNA samples with fluorescent dyes using different methods. Some RNA labeling methods also amplify RNA samples at the same time. At the hybridization step, each labeled nucleotide acid molecule (mRNA, cRNA, or cDNA) in targets hybridizes with a probe complementary to its sequence on the array surface and forms a probe-target hybrid. Depending on the type of DNA microarray, two targets from different sources can hybridize on one array or on two separated arrays. Fluorescent dyes on each probe-target hybrid glow when stimulating the array surface with light. The intensity of emission from each hybrid is proportional to the relative abundance of an expressed transcript. Florescent emissions are captured by optical devices, converted to digital signal, and saved in an image file. Signal intensities stored in image files can be transformed to numerical intensity values for every probe on the array by appropriate algorithms.

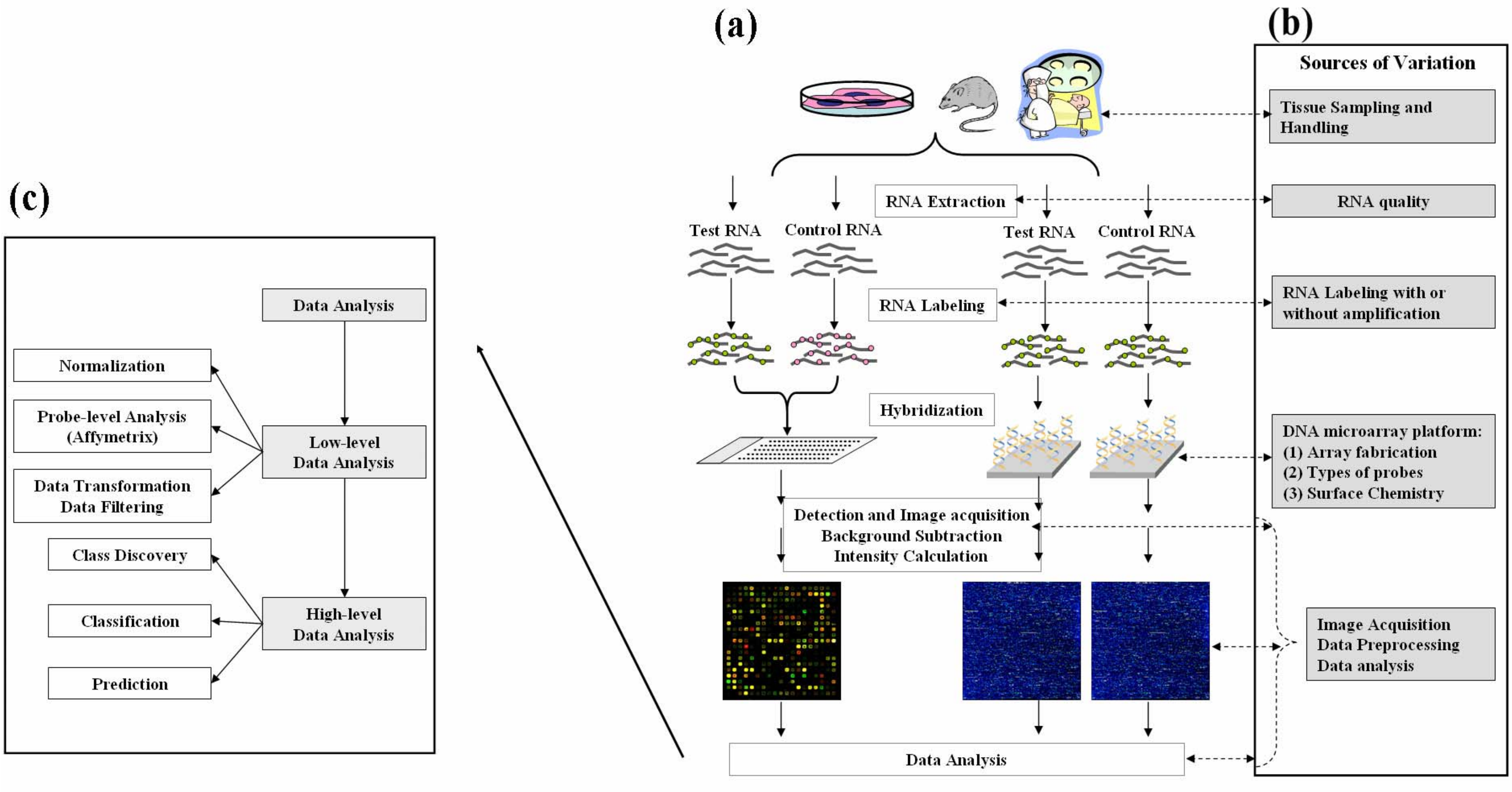


Figure 1.2 Schematic of the experimental procedures, data analysis and sources of variation in gene expression profiling using DNA microarrays.

(a) experimental procedure starting from identifying the RNA sources and all the way to data analysis, **(b)** the possible sources of variation identified at each step of the experimental procedure, **(c)** data analysis including low-level and high level analysis.

1.2.2. RNA quality

The primary target for gene expression analysis is mRNA derived from cell cultures or tissue samples. The quality of RNA from these samples will determine the quality of the ultimate gene expression data from DNA microarray hybridization. Alternatively, the quality of RNA samples is determined by multiple factors in the tissue acquisition and storage procedures discussed in Section 1.5.3.1 and 1.5.3.2. There are several factors that affect RNA quality in the target preparation and labeling process worthy of discussion here. RNA is not stable and RNAases exist in abundance in cells and in laboratory environments, which may cause the degradation of RNA samples. Therefore, RNA samples should be handled with great caution. RNA samples should be purified before target preparation and labeling process to remove proteins, DNAs, etc… In addition, phenol which is a reagent used commonly for RNA purification should be removed as it interferes the labeling efficiency of cyanine fluorescent dyes[57].

RNA purity is usually checked by a UV spectrometer. RNA with good purity should have an OD 260/OD 280 ratio above 1.8. 260nm and 280nm are the wavelengths at which the absorbance of the RNA recorded. If RNA is contaminated with proteins, the ratio will be lower than 1.8. RNA quality can be accessed by electrophoresis on an agarose gel and recently by a Bioanalyzer from Agilent (Foster City, CA). Good-quality RNA on a Bioanalyzer output should have smooth baselines with no increase of small molecular weight fragments. The 28s and 18s (for eukaryotic RNA samples) rRNA ratios should be between 1.5 and 2.0.

1.2.3. Target preparation and labeling

This section describes the approaches to prepare targets from purified, high-quality RNA samples for DNA microarray hybridization. After hybridization of targets with their corresponding probes on a DNA microarray, a reporting system is needed to signal the detection of a target and the intensity of the hybridization. Therefore, like conventional Northern and Southern blots, targets are labeled for signal detection purpose. According to Schena [3] there are two major approaches for target labeling: (1) direct labeling where labeling tags are attached in a covalent manner directly to the target molecule using a enzymatic or chemical means; (2) indirect labeling in which labeling tags are attached in non-covalent and indirect way to the target molecule using dendrimers, antibodies or some other reagent. Because of recent improvement in labeling techniques, mRNA can also be directly labeled without “transferred” to cDNA or cRNA[58, 59].

For DNA microarray based experiments, the most popularly used labeling reagents are various kinds of fluorescent reagents although there are other types of labeling reagents used. Fluorescent reagents can be used for both direct and indirect labeling. Most commonly used fluorescent reagents are cyanines including Cy3 and Cy5, and phycoerythrin (PE)[3].

In addition to labeling, target mRNA may be converted to cDNAs or cRNA for several reasons: (1) labeling can be introduced during the conversion, (2) cDNA is more stable than mRNA and therefore it is easy to handle and store than mRNA, (3) the process of obtaining cDNA and cRNA involves mRNA amplification that provides means to obtain enough targets from minute amounts of mRNA.

To generate good quality array data, a typical microarray experiment requires 5~20 micrograms of labeled targets which correspond to milligrams of tissues or approximately $10^{6\sim}10^{7}$ cells from cell cultures[2, 4]. This requirement limits the use of microarray based gene expression analysis as most clinical tissue samples, for example biopsies and tissues from laser capture microdissection (LCM), are too small to provide enough starting materials. Therefore, enzymatic amplification methods have been applied to generate abundant targets for hybridization from limited amount of starting material[60-64].

In the following sections, target labeling and/or target amplification methods will be discussed. Table 1.1 summarizes these methods based on the amount of total RNA needed, signal amplification or target amplification, popular dyes used, array platform used, time and work intensity. For a graphical illustration of each method please refer to Richter et al.[65].

Table 1.1 Comparison of target labeling methods.

	Reverse Transcription	RT with aminoallyl	T7 RNA polymerase based *in vitro* transcription	SMART PCR	Dendrimer	TSA	Direct labeling of mRNA (1) by Cole et al. (2) by Gupta et al.
Start amount of total RNA	10~15 μg	10~15 μg	0.01~15μg	0.05~1 μg	1μg	0.5~2μg	10~15 μg
Sequence sense	Anti-sense						Sense
Probe sense	Sense						Antisense
Signal amplification	No		Yes. Through passive target amplification	Yes. Through passive target amplification	Yes.		No
Target amplification					No		Yes
Array platform used	Affymetrix, Oligo, cDNA	Oligo, cDNA	Affymetrix, Oligo, cDNA	cDNA arrays	cDNA arrays	cDNA arrays	Affymetrix, cDNA array
Time (hours)	Several hours (4)	Several hours (9)	Several days	Several hours (4)	Several hours (6)	Several hours (9)	2 hours for (1) under 1hour for (2)
Work intensity	Low	medium	High	low	low	High	low

RT: reverse transcription;
TSA: tyramide signal amplification;

1.2.3.1. Labeling with reverse transcription (RT)

The basic RT approach is as follows. An oligo-dT primer is used to hybridize with mRNA which serves as the template for RT. cDNA is reversely transcribed from mRNA by incorporating deoxynucleotides base-by-base, some of which are covalently linked by a fluorescent reagent, with reverse transcriptase[3, 66]. Therefore, in the cDNA sequences produced there are nucleotides labeled with fluorescent dyes. After RT, the mRNA or total RNA templates are degraded and cDNAs are then purified. Both mRNA and total RNA can be used for RT reaction, but only the mRNA molecules in a total RNA sample is reverse transcribed because of the primer. Random primers are used when reverse transcribing prokaryotic RNA samples or RNA without poly-A tails. Reverse transcription of mRNA was the first labeling method used in DNA microarray based gene expression analysis[2, 4]. It is also a simple and effective approach. Furthermore, many different types of fluorescent reagents can be used for labeling with reverse transcription.

Aminoallyl nucleotide analogs have been developed recently for direct labeling using reverse transcription [3, 28, 67]. An aminoallyl nucleotide has aliphatic primary amine group which is subsequently labeled with an amine-reactive fluorescent dye. The labeling is accomplished in two steps. First, the aminoallyl nucleotide analogs are incorporated through reverse transcription. Then the amine-reactive fluorescent dye reacts with aminoallyl group of nucleotide incorporated to form covalent bounds. Compared to fluorescently labeled nucleotides, aminoallyl nucleotide analogs have higher incorporation rates because of its small size and therefore more effective labeling can be achieved.

1.2.3.2. Labeling with T7 RNA polymerase based *in vitro* transcription

The T_7 RNA polymerase based *in vitro* linear amplification assay makes single stranded cRNAs that is antisense to the corresponding mRNA molecule by a multi-step enzymatic reaction [60, 61, 63]. An RNA polymerase promoter is integrated into 5' end of each cDNA molecule by reversely transcribing mRNAs with a chimeric primer composed of oligo (dT) with the bacteriophage T_7 RNA polymerase promoter. The 2nd strand of cDNA is then synthesized using the 1st as a template. Once double stranded, the T_7 RNA polymerase promoter region becomes functional. The ds-cDNA molecules subsequently serve as the template for in vitro transcription of the single stranded antisense complementary RNA molecules by T7 RNA polymerase. Biotinylated labeled NTP molecules are introduced at this step. The degree of amplification depends upon the enzyme concentration and incubation time[61]. In a typical application of this method, more than 30 μg of cRNA can be generated from as little as 1μg of total RNA using an overnight incubation. When only minute amounts (less than 100ng) of total RNA are available, this method is modified to include an additional cycle of reverse transcription and in vitro transcription after the first in vitro transcription reaction in order to produce enough cRNA targets[60, 63]. Aminoallyl nucleotide analogs have also been used recently for direct labeling with the T_7 RNA polymerase based in vitro amplification[68].

Fidelity and linearity of the approach have been well documented (Section 1.5.4). This approach provides a solution for the problem of insufficient amount of target mRNAs for microarray hybridization. However, this approach is quite time consuming and several studies have shown systematic biases introduced by this approach[64, 69, 70]. More details on comparisons studies of labeling approaches will be summarized and discussed in Section 1.5.4. The variations introduced by this method have also been investigated thoroughly in Chapter IV.

1.2.3.3. Labeling with SMART™ PCR amplification

Polymerase chain reaction provides a way to amplify bulk amount of nucleic acids from minute amount of starting material. However, the nature of PCR limits its use in target preparation for microarray based gene expression experiments. PCR amplification is exponential where short templates are amplified more efficiently than long templates and therefore will distort the abundance and complexity in the mRNA population studied[64]. Nevertheless, there are several studies used a modified version of PCR amplification, SMART™ PCR or template switch PCR (TS-PCR), for target amplification and labeling in microarray gene expression experiments[62, 64, 71]. Based on the report from Petalitis et al.[62], SMART™ PCR provides better linearity compared with traditional PCR and increase the sensitivity of microarray experiments by allowing the detection of more transcripts which are below the detection concentration using direct labeling. Saghizadeh et al.[71] showed SMART™ PCR and IVT amplification are comparable in reproducibility and reliability by comparing gene expression data from the two approaches to data form total RNA and mRNA. However, a comparison study of SMART™ PCR and the T_7 based amplification method by Puskás et al.[64] demonstrated that gene expression data from the T_7 based method had better correlation with nonamplified (labeled by RT reaction) data than data from SMART™ PCR even though, both approaches gave highly reproducible data. Nonetheless, PCR amplification methods have not been accepted as a safe, reliable target labeling and amplification approach.

1.2.3.4. Tyramide signal amplification

Indirect labeling using tyramide signal amplification (TSA) employs antibody-antigen binding and enzymatic reactions to label array spots of target-probe hybrids. Labeling with TSA labels target-probe hybrid and therefore labeling event takes place after hybridization. In addition, TSA labeling generates highly reactive fluorescent dyes which then couples with the array surface. Two components are critical to this assay: hapten (usually biotin) modified nucleotides and horseradish peroxidase (HRP) conjugated antibody (streptavidin). cDNA are incorporated with hapten modified nucleotides by reverse transcription of mRNA/total RNA and then hybridized on the array. After hybridization, the array is incubated with HRP conjugated antibodies, which bind to haptens on the target-probe hybrids. The binding of antigen and hapten brings HRP close to the array surface. The array is then incubated with hydrogen peroxide (H_2O_2) which is used by HRP to oxidize tyramide linked fluorescent reagents such as Cy5-tyramide. Oxidized tyramide fluorescent reagents are highly reactive and can rapidly attach to the array surface. Therefore, at only the spots where hybridization takes place, fluorescent dyes will be incorporated[3, 65, 72].

TSA approach amplifies fluorescent signal rather than the target. The advantage of using TSA compared to using the reverse transcription method is that it can provide hundred fold signal amplification and therefore minimizing the usage of precious RNA samples[3, 72]. One drawback of this protocol is that it is a little time consuming compared to other methods[3].

Biotin labeling

It is worthwhile to reveal the mechanism of biotin labeling approach which is adopted by several oligonucleotide platforms such as Affymetrix GeneChip® arrays[2] and CodeLink BioArray [25, 73] for gene expression from GE Healthcare. The biotin labeling method is similar to TSA as it also uses antigen and antibody binding (biotin and streptavidin) to link fluorescent dyes to the target-probe hybrids but there is no enzymatic reaction needed for the biotin labeling approach. Streptavidin molecules are coupled with fluorescent dyes, phycoerythrin (PE) for

GeneChip® arrays and Cy5 for CodeLink bioarrays, and therefore the labeling is done once the antibody and antigen binds.

1.2.3.5. Labeling with dendrimer

A dendrimer is a complex nucleic acid structure created by specific annealing of multiple nucleotide oligomers together to form a highly-branched structure[65, 74-76]. When use for labeling in microarray experiments, some branches can be labeled with fluorescent reagents, other branches tagged with a specific sequence tag.

The indirect labeling approach using dendrimer works as follows. At reverse transcription, a chimeric primer is use, which contains both the poly-dT primer sequence and a small piece of sequence called “capture sequence”. This capture sequence is integrated to the 5’ end of cDNA molecule. (The capture sequence is complementary to the aforementioned sequence tag in some dendrimer branches.) cDNA targets are then hybridized on the array. cDNA are incorporated with hapten modified nucleotides by reverse transcription of mRNA/total RNA and then hybridized on the array. After hybridization, the array is incubated with the dendrimer. The sequence tag on a dendrimer hybridize with the capture sequence at each target-probe hybrid and the fluorescent labels on the dendrimer will be used to report the target-probe hybridization[3, 65, 76].

The advantage of dendrimer approach is that each dendrimer can carry hundreds (up to 350) of fluorescent labels. Small amount of target molecules (as low as 1 μg) at hybridization will provide enough signal for detection with high signal-to-noise ratio. Labeling process does not depend on the incorporation of fluorescent dNTP in a RT or IVT reaction and therefore the labeling can be robust[3, 65, 76].

1.2.3.6. Direct labeling of RNA

There are labeling approaches developed to label RNA molecules directly without amplification or conversion in the second example. Two methods are worthy of discussion.

Direct labeling of RNA with multiple biotins has recently been proposed by Cole et al. from Affymetrix Inc. (Santa Clara, CA)[58]. This approach uses T4 RNA ligase to attach a 3'-biotinylated donor molecule to the target RNA[58]. mRNA targets are reverse transcribe to cDNA and then converted to cRNA by in vitro transcription without amplification. cRNA are then fragmented to small pieces, 50~200bp, and dephosphorylated to expose 3'-hydroxyl groups. T4 RNA ligase is then used to catalyze the addition of the biotinylated donor molecules to 3'-end of the fragmented cRNAs. This approach labels fragmented cRNAs uniformly and therefore can avoid any bias from sequence dependent incorporation. Gene expression data from this method has 90% agreement with data from T_7 based labeling approach when studying gene expression in AML vs. ALL indicating its acceptable sensitivity for detecting differentially expressed transcripts[58].

Direct labeling of RNA with platinum-linked cyanine dyes was reported by Gupta et al.[59]. In this approach, the mRNA molecules are labeled in the total RNA by a single-step non-enzymatic reaction. Platinum is attached with cyanine fluorescent reagents. The platinum reagent can react with the N_7 of guanine (G) residues in a RNA sequence to form a stable coordinate bond[59]. The stability of the RNA molecule and its hybridization ability is not be affected by labeling.[59] By comparing to gene expression data from labeling with regular reverse transcription approach, Gupta *et al*.[59] demonstrated the high precision and low error for gene expression analysis.

The advantage of direct labeling of RNA is that it does not require any amplification and has only minor enzymatic manipulations. The integrity of the RNA sample is maximally preserved and the labeling procedure is much simplified. In addition, the direct labeling of RNA is quite fast. On the other hand, the direct labeling of RNA requires large amount of RNA and therefore may not be realistic for studies using low volume or rare tissue samples[58, 59]. One important note is that targets from direct labeling of RNA approaches are sense sequences in reference to mRNA, i.e. the same sequence as mRNA, and therefore the DNA microarrays need to be constructed to contain antisense probes[58, 59].

1.2.4. Hybridization, detection and image acquisition

There are two major ways to perform hybridization experiments (Figure 1.2a). A target can hybridize on one DNA microarray alone or two targets of different origin can co-hybridize on the same DNA microarrays. When using co-hybridization, targets are typically prepared from RNA taken from either cell cultures under different conditions or two different tissue samples and the same transcripts/genes from the two targets will compete for probe binding on the DNA microarray. Therefore, targets need to be labeled with two different fluorescent reagents so that binding of transcripts from each target can be distinguished at each spot. For a transcript/gene, depending on the amount of the two targets, the hybridized spot will show mixed color. Co-hybridization usually is used on spotted arrays; the most popular pair of fluorescent dyes is Cy3 and Cy5. On the other hand, if hybridization is done one target per DNA microarray, targets can be labeled and prepared with an appropriate approach and the comparison of two samples is conducted afterwards using intensity data from the two DNA arrays.

Hybridization of targets with probes on DNA microarrays typically takes place at certain temperature and requires a certain amount of time to ensure hybridization reactions approach equilibrium. A closed environment is needed to retain the hybridization solutions on the array surface. Affymetrix GeneChip® arrays use a closed plastic chamber; other DNA microarrays made on microscopic glass may need a lid with sealed edges to cover the array surface. Depending on the types of labeling approaches, extra work may be needed after hybridization process. For example, if targets were prepared using Dendrimer or TSA approaches, the actual labeling takes place after hybridization[3, 66, 72, 74]. Once all steps have been completed, DNA microarrays are subject to scanning and image acquisition.

Hybridization signals are generated by using light to stimulate the emission of the fluorescent dyes. Light from a lamp with certain wavelength shires on the array surface and causes the fluorescent dyes at each feature/spot to emit. The amount of emission is determined by the amount of fluorescent dyes bound, which is correlated with the amount of targets-probe hybrids at the site. If co-hybridization takes place, two wavelengths are needed to stimulate emissions of the two different fluorescent dyes. The emission is passed through emission filters and captured by a charge-coupled devise (CCD) camera. A scanner walks through the entire array surface with precise positioning to capture the signal feature-by-feature. Signals will be stored in image files. Therefore, image files are the first set/level of data generated from microarray experiments. For Affymetrix GeneChip® arrays, the image files have ".dat" extension. For cDNA arrays, the images files are usually in TIFF format.

1.2.5. Background subtraction and intensity calculation

Images must be analyzed to identify the arrayed spots and to measure the relative fluorescent intensities for each element. Most scanners or array platforms provide software packages to handle image processing. Image processing involves three steps. First each spots or features must be identified and distinguished from noises and backgrounds on the array which may be due to target precipitation contaminants (dust, etc.), or hybridization artifacts. The background needs to be calculated locally and then eliminated from each spot or element. Followed by determination of local background, the background-subtracted hybridization intensity for each spot or element must be calculated.

For cDNA microarrays, there are two major ways regarding the calculation of intensities: either using the mean or the median intensity for each spot[3, 5, 77]. The intensity values will be used for normalization and expression ratio calculation that will be discussed in Section 1.3.1.

For Affymetrix GeneChip® arrays, the fluorescent intensity at each probe cell is calculated in several steps. First, a grid is applied to the array image. This grid divides the image into many probe cells. Each probe cell covers an area, measured by pixels, on the image file. Then, pixels at borders are eliminated from intensity calculation as background. The distributions of intensities from remaining pixels in a probe cell are plotted and the intensity value at 75% is used to represent the average intensity of this probe cell. Average intensity is calculated for each probe cell in the image. The resulted average intensities are stored in a text file with ".cel" extension. This file is then used for probe-level analysis and normalization to generate gene expression value for each gene on the array [78].

1.3.Low-level data analysis

Data analysis includes both low-level data analysis and high-level data analysis (Figure 1.2c) When analyzing a microarray data set, low-level analysis is performed first. The purpose of low-level data analysis or normalization is to adjust for any bias which arises from variation in the microarray technology rather than from biological differences between the RNA samples or the printed probes.

1.3.1. Low-level data analysis for cDNA microarrays

For cDNA microarrays, after image processing, it is necessary to normalize the relative fluorescent intensities in each of the two channels and normalize across arrays. Normalization adjusts for variations in labeling and detection efficiencies for the fluorescent labels and for possible variations in initial RNA. It is well known that Cy3 and Cy5 have different incorporation rates[79]. Three normalization methods are popularly used[3, 77, 79]. All methods assume that most of genes in the array, some subsets of genes, or a set of exogenous genes spiked into the RNA before labeling should have an average expression ratio equal to one.

The first normalization method simply uses the total measured fluorescence intensity on an array. The assumptions are: (1) equal amounts of two samples were hybridized on the array; (2) the overall intensity across all spots in the array should be equal for two channels. Under these assumptions, a normalization factor can be calculated and then used to scale the intensity for each spot in the array.

The second approach uses linear regression techniques. For closely related samples, the expression levels of many genes are assumed constant across samples. Therefore, a scatter plot of measured intensities from Cy3 and Cy5 should have a theoretical slope of one. Under this assumption, regression techniques can be used to compute the real slope and adjust it to one.

Usually, in many experiments, the measurements are not linear. Under this circumstance, local regression techniques can be used such as LOWESS regression (Locally Weighted Scatter plot Smoothing regression)[79, 80].

The last approach, developed by Chen et al.[81], relies on the assumption that intensities of a subset of house keeping genes have constant mean values and standard deviation independent of samples. The measured Cy5 to Cy3 ratio for these genes can be modeled and adjusted to one. In Chen et al. study[81], the authors developed an iterative procedure to perform normalization using this approach.

After normalization, the data for each gene is reported as expression ratio of the two samples hybridized on the same cDNA array and ratios are usually in log scale. This ratio for each gene is simply calculated by dividing the normalized intensity from one channel with its normalized value from the other channel. Expression ratios will be used for high-level analysis for class discovery, class comparison, or cancer classification and/or predication.

1.3.2. Low-level data analysis for Affymetrix GeneChip® arrays

1.3.2.1. Probe level analysis

As discussed in Section 1.1.2.3, Affymetrix GeneChip® arrays for gene expression apply a unique probe design. A transcript is represented by a set of probe pairs (11 pairs or 20 pairs) and each pair of probes contains a perfect match (PM) probe and a mismatch (MM) probe. Both PM and MM probes are 25mers oligonucleotides. The PM probes complement exactly to the corresponding transcript sequences; the MM probe has a mismatch nucleotide at the 13th position. When measuring gene expression, each probe will give an intensity measurement corresponding to the hybridization between the probe itself and the targeted sequences in the hybridization solutions. Therefore, analysis of gene expression data from Affymetrix

GeneChip® arrays should start at the probe level in order to yield a single numerical value for a transcript by summarizing the intensity values from all probes in the probe set for this transcript. There are three algorithms used popularly to perform probe level analysis of expression results from Affymetrix GeneChip® arrays. The following paragraphs will summarize these three methods.

Affymetrix Statistical Algorithm

Affymetrix Statistical Algorithm is embedded into the GeneChip® Operating Software (GCOS) package. Early versions of the package are called Microarray Suite, MAS. Version 5.0 (MAS 5.0) also employs the Affymetrix Statistical Algorithm for data analysis. Version 4.0 (MAS 4.0), however, applied empirical analysis algorithms which were proved to perform inferiorly compared with the Statistical Algorithm discussed below. GCOS provides automated operation and control of instruments (fluidics and scanner), management of experiments and sample information, data acquisition, and data analysis. The Affymetrix Statistical Algorithm is the component in this software package that performs data analysis[82, 83]. This algorithm provides two types of analysis, 'single array analysis' and 'comparison analysis'.

Single array analysis

For the single array analysis, there are two components, the detection algorithm, and the signal algorithm. The detection algorithm tells if a transcript is present or not in the hybridization solution. Using probe pair intensity values (which is stored in the ".cel" files) in a probe set it generates a detection p-value and assigns a detection call to the probe set. Detection calls can be "Present", "Absent" or "Marginal"[82, 83]. The signal algorithm in single array analysis is used to derive numerical expression intensity values from probe pairs in a probe set. Signal is calculated using the One-Step Turkey's Biweight Estimate. Briefly, for each probe pair, the intensity values of MM probe or an estimated value, if MM probe has higher intensity value than PM probe, is subtracted from the intensity value of the PM probe. The adjusted intensity of PM probe is log transformed and then assigned a weight based on the differences of this value to the median of all probe pairs in the probe set. For example, if the log transformed, adjusted intensity of PM probe in a probe pair is equal or very close to the median of all log transformed and adjusted intensities from all probe pairs in the probe set, it will be assigned a high weight. The mean of the weighted intensity values for a probe set is determined and becomes the quantative expression intensity value for the probe set[82, 83]. For single array analysis, probe-level analysis is performed first then data is normalized using the method discussed in section 1.3.2.2.

Comparison analysis

Comparison analysis compares expression from two arrays, hybridized with two distinct samples usually experiment and control, to detect changes in gene expression. Similar to single array analysis, the comparison analysis also has two components, a 'change algorithm' and a 'signal log ratio algorithm'[82, 83]. Before results from two arrays are compared, they need to be normalized or scaled to correct variations between the two arrays. Normalization and scaling will be discussed in 1.3.2.2. For the discussion of comparison analysis, assume results from the two arrays have been normalized or scaled.

The 'change algorithm' detects the ratio of gene expression for each probe set between the experiment array and the baseline array. It also calculates a Change p-value. A change call is assigned to each probe set after comparison based on the Change p-value. The change calls can be "Increase", "Marginal Increase", "No Change", "Decrease", or "Marginal Decrease"[82, 83]. The "signal log ratio" algorithm calculates ratios of gene expression between experiment and control arrays. The signal log ratio represents the level and direction of the change of a probe set.

To get this ratio, the log ratio between the two arrays is first calculated for each probe pair. Once log ratios for all probe pairs in a probe set are gathered, the signal log ratio is computed using a one-step Tukey's Biweight method by taking a mean of these log ratios[82, 83].

Model-based Expression Analysis

The model-based expression analysis algorithm calculates Model-based expression indexes (MBEI) for each probe set based on probe level intensity values across multiple Affymetrix GeneChip® gene expression arrays[84, 85]. It has been implemented in the DNA-Chip Analyzer (dChip) software package[86]. Model-based expression analysis performs normalization first and then calculates the expression intensity values. The normalization method used in this model will be discussed in section 1.3.2.2. For the following discussion, please assume normalization has been performed.

Model-based expression analysis builds a model to estimate the expression of a gene in each sample of a multiple-sample study based on the responses of probe intensity values on the arrays to gene expression changes across samples. It assumes that each probe, PM or MM probes, in a probe pair will have different sensitivity in responding to the expression change of the gene corresponding to the probe set. Therefore, a probe-sensitivity index is estimated for a probe across all arrays in the study. The (PM-MM) difference of each probe in an array is the product of MBEI in the array and the probe-sensitivity index of the probe plus an error term. Here the (PM-MM) deference is known for each probe pair and the probe-sensitivity index can be estimated by surveying the intensity value of the probe across all arrays in the analysis. As a result, the MBEI can be calculated and it is reported as the intensity for that probe[85, 86].

In the newer version of the MBEI algorithm, the authors proposed a revised model that calculates the MBEI using only the intensity value of the PM probes. The reason for using only

PM probes is that some MM probes are not sensitive enough to the changes of expression of a transcripts between samples[84]. The inclusion of intensity of MM probe, therefore, may impair the MBEI on faithfully representing the expression level of a transcript in each sample. The PM-only model[84, 86] is similar to the PM-MM model but it changes the (PM-MM) difference term in the algorithm with intensity of PM probe.

Compared to the MAS5.0 algorithm, MBEI reduces the variability of expression intensity estimate of the low expressor or rare transcripts. The PM-only model also provides a better estimation compared to either (PM-MM) model or MAS5.0[84, 85, 87].

Robust Multi-array Average (RMA)

Robust multiple average (RMA) [87, 88] is implemented in the Bioconductor software package [89]. RMA also requires normalized data before probe-level analysis using quantile normalization discussed in section 1.3.2.2. RMA estimates the expression intensity values for each probe on an array using a log scale linear additive model. The model calculates the log transformed PM intensity values that are normalized and background corrected. The expression intensity value of a probe set is obtained by fitting a linear model with the intensity values from each the transformed PM probes (background corrected and normalized)[87, 88].

In a comparison of the three probe-level analysis methods discussed above by Irizarray et al.[87], results demonstrated that RMA outperformed of both MAS5.0 and dChip on the precision of estimating gene expression intensities for low expressor by effectively reducing variations within replicate arrays, the consistency of fold change estimation and on detection of differentially expressed genes (increased sensitivity and specificity of the detection).

1.3.2.2. Data normalization

Probe-level analysis algorithms and/or software packages usually have normalization strategies implemented as well.

Normalization methods used GCOS software

GCOS software package provides two normalization procedures, normalization, and scaling, for the gene expression data Affymetrix GeneChip® arrays. Both methods maneuver either the average intensity value of all transcripts on each array or the average intensity values from selected probe sets/transcripts on the array. If all probe sets are used, the normalization is considered global. The assumption for global normalization is that most of genes in a transcriptome will not have changes in expression in both control and experiment samples, or under disease and normal states. However, under certain circumstances, for example treating cell culture with a drug, transcription levels of many, if not most, genes in a transcriptome will change. Therefore, only the group of genes known with no differential expression can be used in the calculation of average intensity value in normalization procedure. By doing so, the real gene expression change due to drug treatment will be preserved and only the experimental variations are eliminated by normalization. In addition, both methods are considered linear as the average intensity values of all arrays are normalized by multiplying a factor. The factor is calculated by comparing the average intensity value to either an arbitrary number or the average intensity value of a baseline array.

When scaling, a numerical value, which is user defined and adjustable, is set arbitrarily as the target intensity value. The average intensity value of each array is set to this target intensity value by multiplying it with a factor, called scaling factor. If the average intensity value is larger than the targeted value, the scaling factor for this array is less than 1.0. Conversely, scaling factor is larger than 1.0. Scaling makes it possible the comparison of multiple arrays in a study since scaling, presumably, removes possible technical variations across arrays. Scaling can be used independent of the comparison analysis. Therefore, in a multi-array study, even though there may not be a baseline array, the arrays can still be compared for changes in gene expression after scaling[82, 83, 90].

Normalization, on the other hand, can only be used in the comparison analysis (mentioned in Section 1.3.2.1) where gene expression between an experimental array and a baseline array are compared. At normalization step in the comparison analysis, the average intensity value of the experimental array is brought up or down to the average intensity value of the baseline array by multiplying the average intensity value of the experimental array with a factor called normalization factor. If the baseline array is changed, the normalization factor of an experimental array changes as well[82, 83, 90].

Normalization method used dChip – Invariant set normalization

For model-based expression analysis, probe intensities on an array need to be normalized before calculating MBEI. The normalization strategy used here is called "invariant set normalization" [84, 86]. This method normalizes the average intensity value of each array in a study (except the baseline array) to the average intensity value of the baseline array. The baseline array usually is the array with the median average intensity value across all arrays in the study. Alternatively, if a baseline experiment/sample is included in the study, the array from the baseline experiment/sample is used.

This normalization strategy depends on genes of which expression intensities do not change across arrays (i.e. not differentially expressed). However, it is usually not possible to identify such group of genes. The assumption for this method is that probes of a gene with no change in expression between an array and the baseline array will have similar intensity ranks.

Therefore, in this normalization method, for each array, intensity values of probes (only PM probes are used) are ranked. Then the rank from an array is compared to the rank from the baseline array. The probes that have the similar rank on both arrays are considered to belonging to the genes with no differential expression across the two arrays. An iterative approach is used to compare the ranks and select the set of invariant probes across all arrays in the study in the end. Each array is normalized to the baseline array based on the comparison of intensity values of probes in the invariant set across the two arrays. For each study, the invariant set of probes is different[84, 86].

Normalization method used in RMA -- Quantile normalization

The normalization method used in RMA (Robust Multi-array Average, Section 1.3.2.1) is called quantile normalization. Performing quantile normalization makes the distribution of intensity values of all probes the same for all arrays in an analysis. In order to do so, first, data from all probes are projected to a high-dimensional space where each array is represented as a column of intensity values. If there are n arrays in the analysis, after the transformation, there will be n columns of data. Intensity values in a column are then sorted from small to large. Average is taken for all intensity values in a row and this average is used to replace the original value in each cell (i.e. the row in each column). After replacing each cell with the row average, the column is sorted again from small to large. The value in each cell becomes the normalized intensity value for a particular probe in an array[87, 91].

Bolstad et al.[91] compared the three normalization methods aforementioned. Their results show that the quantile normalization method outperformed both the scaling normalization (a linear normalization) and the invariant set normalization (non-linear normalization) in terms of the ability to reduce variations of a probe set across all arrays in an analysis and speed.

1.4.High-level data analysis

1.4.1. High-level analysis overview

Many data analysis methods have been developed or adapted from other fields of research and then applied to DNA microarray technology for gene expression profiling. These methods are usually considered high-level analysis approaches in contrast to the low-level data analysis methods. The development and application of an analysis method is determined mainly by the study objective.

Typical research in biological or biomedical field is hypothesis-driven because it investigates the mechanisms of a specific gene by manipulating the experimental conditions. Microarray based gene expression profiling, on the other hand, provides global/comprehensive survey of gene expression in a transcriptome by monitoring the expression levels of tens of thousands of genes simultaneously. Therefore, most microarray based gene expression profiling studies are typically considered descriptive research[92, 93]. Nonetheless, a gene expression profiling study using DNA microarrays will always at least have clear objectives and answer well-defined questions.

In the following paragraphs, analysis methods will be summarized in the context of study objectives. Since there are a large number of analysis methods available and usually only a small number of them have been accepted and widely used in the research community, only those methods will be mentioned in the discussion. Before discussing the specific analysis objective and methods, it is worthy to describe the gene expression data sets from DNA microarray experiments.

1.4.2. Characteristics of microarray gene expression data

As mentioned before, the power of microarray based analysis is to make possible the simultaneous monitoring of tens of thousands of genes, in a massively parallel fashion and across many samples at the same time. Gene expression data resulted from microarray experiments have unique features which are different from results of all other means of gene expression profiling analysis such as Northern Blot, SAGE or RT-PCR. Let us assume there is a study using microarray experiment to survey gene expression of $\boldsymbol{m}$ samples with an array of $\boldsymbol{n}$ probes/probe sets. The number of data points or expression intensity values in the results will become $\boldsymbol{m} \times \boldsymbol{n}$. Such a study may survey gene expression in a number of samples (several hundreds) and a DNA microarray typically contains thousands of probes. Therefore, microarray data always has large volume and high dimensionality. These features make the analysis of gene expression data from DNA microarray experiments with classical statistical method almost unfeasible. In addition, as multiple samples and genes are surveyed at the same time in a microarray experiment, there may be uncover networks or patterns of gene expression which can not be detected by classical analysis methods.

1.4.3. Study objective and high-level analysis methods

1.4.3.1. Class discovery

Class discovery studies aim to discover previously unknown subtypes of samples or specimens using gene expression profiles from microarray experiments. The main idea for this type of studies is that, although some samples or specimens may share similar morphological features, they have distinct patterns of gene expression. Therefore, a global survey of gene expression will help to discover specific patterns, which can be used to identify this new subgroup of samples or specimens. Class discovery is usually combined with class comparison and prediction to perform cancer classification with microarray gene expression profiling results. Many studies in cancer research have been able to identify new subtypes of cancer using microarray gene expression data[33, 94, 95]. For example, Alizadeh et al.[95] conducted studies to characterize gene expressions in B-cell malignancies systematically. In their study, gene expression patterns identified from microarray data were able to distinguish diffuse large B-cell lymphoma into two clinically significant groups that had not been identified previously using classical morphological criteria and/or cellular markers.

Alternatively, class discovery can also be used to discover groups of co-expressed genes and some of them may be novel genes or expression sequence tags with no known function. By grouping unknown genes with known genes, inferences can be possibly made on the functions and family of a novel gene based on its expression pattern.

Data analysis involved in class discovery is not supervised with predefined class memberships or gene function groups. The resultant clusters or groups of samples or genes are derived based solely on gene expression data with no prior information about the samples. Therefore, methods used in class discovery are "unsupervised"[96, 97]. The most popularly used methods in class discovery are clustering methods including hierarchical clustering,[96, 98] K-means clustering,[96, 99] self-organizing maps[96, 100], etc. In addition, methods for data visualization using multidimensional scaling, such as principle component analysis[96], have been used in class discovery.

Clustering algorithms group similar objects (tissue specimens or genes) by calculating the similarity or distance (dissimilarity) between objects and grouping similar objects together[96]. It is useful to partition genes or tissue specimens into clusters based solely on microarray gene expression results. On the other hand, clustering algorithms can find clusters even in random data sets. Therefore, it is critical to validate the clusters resulting from clustering analysis[96, 97, 101]. Many cluster validation approaches are available[96, 101] but most of them are not trivial to implement and use. When clustering is used in class discovery, researchers should be aware of its limitations and be cautious on any conclusive statements based on clustering results, unless appropriate validation is performed.

1.4.3.2. Class comparison, class prediction and prognostic prediction

Class comparison studies compare gene expression profiles from different classes of cancer specimens and identify gene expression patterns uniquely associated with a particular class or subclass of a certain type of cancer. Class comparison involves two aims. First, gene expression profiles from different classes or subclasses are compared to investigate if there are differences in patterns of gene expression. Secondly, analysis of gene expression is focused on specifying the group of genes differentially expressed between classes if they have distinct gene expression patterns. Most of these studies also attempt to develop a predictor or classifier based on the expression values of selected genes. This last aim, class prediction, may also be listed as a distinct objective. However, most likely, to achieve a possible prediction, studies have to have class comparison completed first.

There are many examples for class comparison studies in the literature. The work from Golub et al.[32] is likely the first publication on class comparison and class prediction. In their study, the authors proposed the classification of human acute leukemia using solely DNA microarray data. They built a class discovery procedure to identify differentially expressed genes between two types of acute leukemia, ALL and AML. Then a predictor is built, using only gene expression results, which can distinguish ALL and AML without prior knowledge of the class membership of a specific sample. This study sets an example for molecular classification using microarray data, which combines class comparison and class prediction to achieve better classification of cancer. Many more studies have followed this work and attempt to molecularly classify of many different types of human cancers[33, 36, 45, 102]. A number studies are also focused on the classification of multiple cancer types using microarray gene expression results[33, 34, 39, 40, 103, 104].

In addition to classification, recent studies focused on linking gene expression patterns from microarray results to the prediction of cancer prognosis and clinical outcomes. This kind of study aims to investigate whether there is a relationship between gene expression and clinical outcomes. Usually a prognostic predictor is built using solely the expression levels of the selected genes to predict the clinical outcome. Studies have focused on predicting outcomes, such as metastasis and patient survival, of many types of human cancers[34-36, 38, 105-108]. By comparing predictions based on classical histologic criteria, predictors built on gene expression results have been proved more powerful in predicting the outcome of the cancer from a patient and therefore will help to improve therapeutics and patient care.

Van't Veer et al.[106] reported the prediction of the clinical outcome of breast cancer using microarray data. The authors were able to identify a 70-gene prognosis signature or poor prognosis predictor of distant metastases of breast cancer with high accuracy even though the primary cancer was lymph node-negative disease, i.e. no lymph node metastasis at the time of diagnosis. Van De Vijver et al.[107] further evaluated this "poor prognosis predictor" on its ability to predict the survival of 295 breast cancer patients with and without lymph node metastases. Their results demonstrated that the predictor/prognosis signature built using only gene expression data performed best on predicting the appearance of distant metastases in first five years after treatment. The predictor was also highly predictive of the risk of distant metastases of patients with or without lymph node metastases at the time of diagnosis. Results from these two consecutive studies proved the predictive power of gene expression signature in clinical outcome studies.

The prerequisites for successful cancer classification and prognosis prediction are comprehensive information related to clinical outcomes including patient demographics, pathological and histological information such as disease stage and grade, treatment, prognosis related information such as lymph node metastases and survival after treatment, etc... Gene expression signatures with high predictive ability can only be identified when informative and sufficient clinical information is provided along with the gene expression profiles from microarray experiments. The collection of cancer tissue specimens and the management of the patient and disease related information requires tremendous effort over a long periods of time and rely on tissue banking informatics tools such as disease-specific databases or comprehensive data warehouses.

If a study involves class comparison, significance tests should be used to identify differentially expressed genes between samples belonging to different classes. Significant tests are usually considered supervised since data analysis is conducted using the known knowledge of sample class memberships. This is in contrast to unsupervised methods such as clustering.
Though many statistical tests are available (and more are being developed), the most popularly used methods are nonparametric tests such as Mann-Whitney rank sum test or multiple testing with controlling for false positives such as Welsh's t-test with the Bonferroni correction. ANOVA can also be used to compare the multiple conditions. Alternatively, Significance Analysis of Microarrays (SAM)[109] and the Signal-to-Noise metric with Permutation[32] allow the estimation of variation in experimental error between replicates and therefore estimate the

false discovery rate at various levels of stringency. In general, the success of significance testing is dependent on the size of the difference one wishes to detect, the variation in the data, and the number of replicates available. Because the variation in the data is a function of expression level (with greater variation at lower expressions), it is difficult to estimate the level of differential expression that will be found to be significant in an experiment. Hence, gene lists from significance tests should be interpreted carefully.

Classification and class prediction also use supervised methods, which can detect patterns of gene expression and build classifiers or predictors. The classifiers or predictors are built based on the expression data of some differentially expressed genes. The selection of differentially expressed genes is called "feature selection". It is the first step toward developing the classifier or predictors. Methods for feature selections can be the methods used for class comparison, which identify significantly expressed genes between samples. Another method, principle component analysis[110], can also be used for feature selection. The popularly used methods for building classifiers or predictors from microarray data are Discriminant Analysis[111], Nearest Neighbor classifiers[38], the Weighted Gene Voting method[32], Support Vector Machines[112, 113], Shrunken Centroids[114], Fishers Linear Discriminant Analysis (FLDA)[111] and Decision Tree Based Methods using Recursive Partitioning[115].

It is difficult to compare the performance of different methods and to identify the best method. Dudoit et al.[111] performed a direct comparison of several methods (Weighted Gene Voting method, Decision-Tree learning method with or without boosting, Nearest Neighbor classifier, standard and diagonal discriminant analysis) on their ability to classify three microarray data sets. Their results showed that the diagonal discriminant analysis and Nearest

Neighbor classifier had the best performance; the standard discriminant analysis performed poorly. For their study, decision-tree learning methods had intermediate performance.

An issue associated with class prediction studies is the validation of the classifiers or predictors. It is usually accomplished by two ways. Cross-validation is performed for almost every study. In doing so, usually a portion of the samples or specimens is left out when building the classifier or class predictor. Subsequently it is used to predict the class membership of the left-out samples. The set of left-out samples is referred to as the ‘test set’; the set of samples used in the classifier is called the ‘training set’. Leave-one-out cross-validation (LOOCV) is a special case of the cross-validation method. In LOOCV, one sample is left out at a time and a classifier is built using the training set of samples (all samples – the left-out sample). This classifier is tested on its ability to predict the class membership of the left-out sample. This process is repeated until each sample is left out once. The performance of classification or class prediction is measured by the overall error rate in LOOCV[32, 96]. It is important to note that, at cross-validation or LOOCV, feature selection is only performed using the training set. The test set is remained unseen by the classifier so that the accuracy on predicting the test set reflects the “true” performance of the classifier[96, 97].

Alternatively, the performance of a classifier or a class predictor can be validated by classifying or predicting the class memberships of samples in another microarray data set. Samples from this new data set should not be used either in feature selection or in developing the classifier. Such a data set can be obtained from the published articles that carried out studies using similar RNA samples as we did in Chapter II[38]. Alternatively, if there are a large number of samples in an analysis, the original data set can be divided into two portions; the larger portion of samples is used to build the classifier and the smaller portion is used to validate its performance[96].

1.5.Variations in gene expression data from DNA microarray experiments

1.5.1. Overview

It is common for multiple groups to conduct similar studies using different types of DNA microarrays and/or clinical tissue specimens. These studies may also ask very similar, if not the same, questions; however, gene expression results from DNA microarray based experiments often give quite different answers due to the lack of control of variation.

For example, the contest data sets for the Critical Assessment of Microarray Data Analysis (CAMDA) 2003 conference came from four studies on lung cancer. The four data sets were created by four research groups from different institutions, Michigan[35], Harvard[33], Stanford[34], and Ontario[36]. Data sets from Harvard and Michigan groups were generated from two generations of Affymetrix GeneChip® arrays; and data sets from Stanford and Ontario used cDNA microarrays. The goals of their studies were: (1) to perform molecular classification and staging of lung cancer specimens for diagnosis purpose, (2) to help predict cancer prognosis and patient survival.

Each study reported either (1) a molecular classification of lung cancer with a panel of differentially expressed genes or (2) linkage of patient survival to gene expression. There was little or no overlap between the differentially expressed gene lists from the four studies. This makes true meta-analysis very challenging. Furthermore, only 4% (2499) of the total transcripts (62029) surveyed by the four DNA microarrays—Harvard 12600, Ontario 19200, Michigan 7129, Stanford 23100—were present on all four of the arrays used in these studies[116]. Other limitations on cross-platform integration were related to inherent differences between types of DNA microarrays (cDNA probes vs. oligonucleotide probes or one channel vs. two channel), tissue sampling and handling, clinical parameters (tumor stage/grade, percentage of tumor cell in the specimen, etc.), demographic differences of the patients etc. Each of these factors may introduce variations and/or systematic biases into gene expression data. Results in Chapter III reported our efforts on attempting to integrate data sets from Harvard and Michigan using two different integration strategies. Although integration can be achieved to certain degree in our results, complete integration is not possible due to the variations in microarray data sets.

Studies on the molecular classification of prostate cancer showed similar observations[37, 38, 40]. Independent groups at University of Pittsburgh[37, 40] and MIT[38] compared prostate cancer to non-cancerous prostate tissue using two versions of the Affymetrix GeneChip® human genome arrays, HG_U95av2 and HG_U95A arrays respectively. The study carried out by the MIT group identified a group of genes strongly correlated with the grade of prostate tumor differentiation measured by Gleason score, and a model built solely on microarray gene expression data was able to predict patient outcome[38]. The study by our laboratories at the University of Pittsburgh, on the other hand, built a gene expression model using 70 differentially expressed genes to predict the aggressive behavior in prostate cancer[40]. Both studies provided strong evidence for the concept that the clinical behavior of prostate cancer are linked to the differences in gene expression patterns which could be detected at the time of diagnosis[38, 40].

On the other hand, although the tumor and non-tumor samples used in the two studies had very similar clinical features and DNA microarray data were generated using very similar arrays, classifiers built based on data from University of Pittsburgh could not classify correctly tissue samples from the MIT study[117]. (Ma, et al. submitted to BMC bioinformatics and please see Chapter II in this thesis for details). Further investigations[37] on the data set from University of Pittsburgh revealed that the non-tumor tissue samples were dissected from regions close to the cancer and therefore carried genetic changes similar to or mimic of morphologic cancer.

Therefore, in this example, tissue sampling likely affected the performance of the classifier built based on data from University of Pittsburgh. This underlines the importance of selecting the correct "baseline" or control tissue in the design of microarray based gene expression experiments[37, 117].

Many more studies in every field of research where DNA microarray is used as the major analysis tool for gene expression face similar challenges for cross-platform comparison. Many other factors, in addition to those aforementioned, influence microarray based gene expression studies profoundly and the resulting gene expression data may suffer from variations or biases caused by these factors (Figure 1.2b). As the awareness of the problem and pitfalls in microarray data has increased dramatically for the last 5 years, cross-platform comparison studies have been used to investigate the types, sources, and extent of variations, and potential methods to minimize these variations. The following section will review these studies and summarize the current knowledge on the sources of variation, which interferes with cross-platform comparison and data integration.

1.5.2. Overall levels of variations indicated by cross-platform comparison studies

Almost all available DNA microarray platforms have been compared with each other in various studies[118-121]. Comparisons were focused on: (1) the concordance of intensity levels detected for each transcript surveyed on an array; (2) the concordance of differential expression ratios between experimental samples and controls or under two conditions; (3) the biological themes identified in gene expression data. Concordance is often measured by the correlation of intensity values or ratios over a group of overlapped transcripts. The group of transcripts overlapped among DNA microarray platforms compared in a study is usually identified by finding transcripts with the same UniGene identifier or by matching the sequences of transcripts targeted by probes. Biological themes are usually assigned to differentially expressed genes from a study by studying their categories in Gene Ontology[122] using software tools such as EASE[123]. In addition to the relative comparisons among arrays, Northern blot analysis, and (quantitative) RT-PCR (qRT-PCR) results of a number of transcripts provide a standard for the assessment of the degree of agreements among various array platforms and further allow calibration of microarray data.

Early comparisons in 2002 and 2003 were rather discouraging as comparison results showed there was lack of concordance between the available microarray platforms that were designed to survey biological relevant patterns[119, 124-126]. The correlations of the intensity values and the ratios/fold changes of the differentially expressed genes across platforms ranged from significantly low (<0.5) to moderate (0.5~0.6)[124, 126]. The subsets of differently expressed genes identified by different platforms had limited overlap[119]. However, it is important to note that for the genes that do overlap, the ratios/fold changes showed good agreement with fold changes from Northern blot analysis or qRT-PCR[124]. On the other hand, these independent validation approaches failed to validate transcripts with disparate expression intensity values across platforms[125].

Over the past several years, cross-platform concordance has improved significantly as both the technology and experiences with the technology advanced. In 2004, Yauk et al.[118] compared six DNA microarray platforms used for gene expression analysis including one with long oligonucleotide arrays, three with short oligonucleotide platforms, and two cDNA array platforms. Results showed rather reasonably high correlations (0.65~0.78) for the pair-wised cross platform comparison of the Affymetrix GeneChip array, the Agilent oligonucleotide array, the CodeLink Uniset I array, and the Agilent cDNA array using either a group of genes common in all platforms or the genes in common on a given pair of platforms. Similar or better levels of concordance have also been reported by other studies[127-129]. Using a two factor ANOVA analysis, results from Larkin et al.[130] in 2005 demonstrated that 90% of the overlapping genes (~5800) between the Affymetrix GeneChip® arrays and TIGR cDNA arrays was little affected by platforms on gene expression intensity values and that these genes had high correlation of intensity ratios across platforms (>0.80). For genes consistent across platforms, qRT-PCR results also showed robust correlations between platforms. While for genes with disparate expression ratio measurements between platforms, qRT-PCR results disagreed with both platforms and provided a third expression profile.

Larger scale, comprehensive studies have been carried out in 2005 on comparing multiple DNA microarray platforms, taking into consideration not only the variations between microarray platforms but also the variability across different laboratories[120, 121].

Irizarry et al.[121] published the first ever platform comparison study across different laboratories in Nature Methods, April 21, 2005. In this study, the authors compared gene expression results from three platforms--Affymetrix GeneChip® arrays, two-color cDNA arrays, and two-color oligonucleotide arrays--using the same RNA samples with a known number of differential expressed genes. Gene expression data were produced from a consortium of ten laboratories from the Washington DC and Baltimore areas with each array platform implemented in at least two laboratories. Their results showed laboratory effects typically influence the precision of gene expression data, and that precision varies across laboratories where the same array platform was used.

In the same issue of Nature Methods, another cross-platform, between-laboratory study was conducted by the Toxicogenomics Research Consortium[120]. In this study, the authors investigated effects from many technical aspects, such as RNA labeling methods and data acquisition, as well as effects from various microarray platforms on the overall agreement of cross-platform comparisons. Gene expression data from a total of twelve DNA microarray platforms and across eight laboratories were compared. For most platforms in the comparison, reproducibility of gene expression data between laboratories was poor but could be dramatically improved by standardizing RNA labeling and hybridization, microarray processing, data acquisition and normalization. This result highlights the importance of standardization in improving the concordances across array platforms. ANOVA analysis of this data demonstrated that the microarray platform is the most prominent source of variability in microarray data and contributes to more than half of the variability observed in the data. Other sources of variations are laboratory, tissue, array replicates, tissue × platform, tissue × laboratory, and dye. Overall, the results showed that good concordances (correlation coefficient >0.90) were achieved across commercial DNA microarray platforms and between laboratories. The best reproducibility came from commercial DNA microarrays with standardized protocols. In summary, within the limits investigated to this point, microarray data sets can be comparable across platforms and between laboratories when known sources of variation have been controlled for or eliminated.

Although reasonably good agreement across platforms can be achieved, it is still far from perfect. In all cross-platform comparison studies, there have been a number of genes which have disparate expression profiles across platforms. For example, a gene can be differentially expressed with a large fold change in expression data from one platform but show no difference across the same biological samples in data from another platform. Northern blot or qRT-PCR may verify none of the results from any of the array platforms but may generate yet another gene expression profile for this gene[121, 130]. This may be due to other sources of variation that have not been identified or more likely it is because transcript sequence information and annotation have not been fully understood or optimized. Either way, the concordance can still be improved.

Regardless of the level of concordance achieved in various comparison studies, there are a number of common findings worthy of discussion. Commercial DNA microarrays typically show lower variances in array replicates and higher sensitivity in detecting differentially expressed genes than custom made/academic center DNA microarrays[118, 120]. Therefore concordances between commercial DNA microarray platforms are better than the agreement between custom made arrays or between custom made arrays and commercial arrays[118, 120, 131]. Concordances across platforms are better when using fold changes/ratios of gene expression instead of intensity measurements[130]. Even when there is low level of concordance across array platforms and/or between laboratories, the biological themes identified in gene expression data from different platforms may have significant agreement with each other[119]. Reproducibility of gene expression data across platforms can be improved by applying "superior" analysis algorithms such as probe level analysis algorithms and normalization techniques[130]. Agreements across platforms can be improved by removing noise in the gene expression data for example by using only transcripts identified as present by preliminary DNA microarray analysis in the initial RNA samples[128] or by applying standardized protocols[120, 121, 130]. Finally, tissue heterogeneity, cell type difference, or biological treatments influence much more significantly the gene expression data than technical variations among different platforms[118, 130, 132].

The most important sources of variation investigated in platform comparisons studies (mentioned above) are summarized as follows (also depicted in Figure 1.2b side-by-side with steps in the experimental procedure Figure 1.2a).

(1) Inherent differences among various platforms such as probe type, probe length, features of probe sequences, probe quality, annotation, and surface substrate,

(2) Tissue sampling and handling,

(3) Technical aspects such as RNA labeling, hybridization, and data handling as demonstrated in the between laboratory comparisons.

The following sections will focus on variations from tissue sampling and handling, and RNA labeling. These discussions will provide background knowledge for Chapters II, III, and IV.

1.5.3. Variation introduced by tissue sampling and tissue handling

The purpose of gene expression analysis is to detect relative changes in gene expression in order to make inferences regarding the underlying biological mechanisms or states. Therefore, gene expression profiles should reflect either the biological or the physiological state of a tissue sample or cell culture or relative differences in gene expression between two tissue samples or cell cultures under different treatments. It is not surprising to discover that biology is the major driving force in cross platform comparisons. For example, results from the study by Yauk et al.[118] showed that, despite the different types of arrays used, gene expression data formed major clusters based on the cell types and tissue origin indicating that biological differences influence gene expression data more significantly than technical variations among different platforms. As mentioned above, Larkin et al.[130] demonstrated that gene expression intensity values of 90% of the overlapping genes (~5800) between the two platforms were not affected by array platforms but rather represent the true states of gene expression in the original RNA samples. A number of comparison studies on RNA preparation and labeling approaches also proved the same idea by demonstrating that RNA amplification and labeling approaches can introduce non-negligible amount of biases/variations but these biases do not significantly impair gene expression profiles[64, 133, 134].

On the other hand, as pointed out above, sampling and tissue handling have significant impact on accuracy and reproducibility of microarray based gene expression analysis. This is of particular importance for microarray based gene expression studies on human diseases such as the breast cancer studies in the CAMDA workshop[33-36] and the prostate tumor classification studies[38, 40] discussed at the beginning of this section. Careful sampling is important because human tissues are very heterogeneous by nature. Tissue handling, on the other hand, ensures the quality of the initial RNA samples obtained from a tissue specimen.

1.5.3.1. Tissue sampling

Human tissues, normal or diseased, are very heterogeneous, comprising many cell types including epithelial cells, stromal cells, muscle cells, nervous tissue, fat cells, immune cells. Different regions of a tissue may also have distinct composition of cell populations (cell type + the number of cells in each type). This results in variation even between very similar tissue samples. This inter-sample heterogeneity (tissue heterogeneity) is in addition to differences between individuals (the inter-patient heterogeneity). For example, individuals may have different single nucleotide polymorphism at a particular site in the sequence of a transcript. In cancer studies, even though tissue samples dissected in a surgery are rich in cancer cells, the population of cancer cells may be “contaminated” by non-cancerous cells. Cancerous tissues removed by surgery typically contain normal appearing tissues to remove completely the cancerous tissues. Therefore, most tumor tissues are “contaminated” with normal appearing tissues. There is also heterogeneity of cellular populations such as multifocality and different degrees of malignancy (grade). Furthermore, as a large tissue specimen usually is dissected into small blocks and each block used for a study is different from all others.[57, 135, 136] Giving this level of heterogeneity, it is quite challenging to obtain consistent tissue samples with a somewhat consistent number of cells affected by the disease or biological state, for gene expression analysis in a study.

One example of the effects of sampling on microarray based gene expression profiling is the study on molecular classification of prostate cancer at University of Pittsburgh[37, 40, 117]. Three types of tissues were collected, cancerous tissues (tumor), normal appearing tissues adjacent to the cancer (adjacent normal), and normal tissues from disease-free (free of any prostate related diseases) organ donors (donor). Each type of tissues produced unique gene expression profiles. Yu et al.[40] reported a cancer field effect because the gene expression profile of the adjacent normal tissues is changed substantially (and, to certain degree, it resembles the profile from cancer). Chandran et al.[37] conducted in depth analysis to compare the gene expression profiles from the three types of tissues with the same data set. Profiles from adjacent normal tissues correlate better with profiles form tumor than those of the donor tissues. The comparison of tumor vs. donor detected many more differentially expressed genes than the comparison of tumor vs. adjacent normal at a similar stringency level (false discovery rate<0.025). These results suggested that the normal appearing tissues adjacent to prostate cancers undergo tumor-like changes in gene expression. The authors also speculated on which type of tissue, normal appearing tissue adjacent to cancer or normal tissue from disease free organ donor, serve as a better baseline in an study for detecting differentially expression genes in prostate cancer. Alternatively, Ma et al.[117] performed classification studies using the same set of data using various machine learning algorithms. Classifiers constructed from profiles of tumor and donor performed better than classifiers built form profiles of tumor and adjacent normal both in leave-one-out validation and in classification of gene expression data sets of prostate cancer from other institutes. These results reinforced the hypothesis of the choice on baseline tissues for prostate cancer studies which should include "true normal" controls.

Interestingly, Stamey et al.[137] conducted a study to investigate what is the best prostate control tissue. Three potential control tissues were collected, peripheral zone, central zone, and benign prostate hyperplasia (BPH). Gene expression profiles were generated from each type of tissues. Profiles were compared to each other for efficiency in detecting present transcripts/genes. Each profile was then compared with expression profile from Gleason grade 4/5 prostate cancer for detecting differential expression genes. Their results showed that there was substantial overlap of the present genes in each profile. However, very little overlap of the differentially expressed genes detected using each "normal" type of tissues as control. Expression profiles of the morphologically normal appearing peripheral zone tissues shared many genes with Gleason grade 4/5 cancer, suggesting a possible field effect similar to that described in the Pittsburgh study. Their results demonstrated the variations introduced by inappropriate sampling and, once again, emphasized the importance of using the right controls for any study on detecting differential gene expression and therefore eliminated from the study.

Several approaches have been proposed and implemented in daily practice to provide good/correct sampling in gene expression analysis. First, pathological evaluations should always be used at the time of collecting tissue samples. Pathological evaluations can help ensure the tissue blocks obtained for a study do contain enough of the tissues (such as tumor tissues) of interest. Pathological evaluations will also provide cell composition information in the tissue samples. For example, pathological evaluation can tell that tumor cells may only be 10% of all cells in a tissue specimen used for gene expression profiling.

Intra- and inter-patient heterogeneity can be diminished by increasing sample size so that appropriate statistical analysis and normalization can be applied to expression profiles. Alternatively, Bakay et al.[132] first demonstrated that intra-patient tissue heterogeneity and differences between individual patients (inter-patient) are the major source of variability in gene expression data when a single DNA microarray platform is used. Then, the authors tested a strategy for eliminating the intra-patient and inter-patient variations by pooling/mixing of patient cRNA samples before hybridization. Pooling was done with cRNA samples from different region of a tissue to help minimizing intra-patient variations. It was also done with cRNA samples from different patients with matched age, gender, disease stage, etc. to diminish inter-patient variations. Results from hybridizations using pooled cRNA samples demonstrated that gene expression profiles from pooled samples were able to detect differential gene expression between target tissue and control with high specificity comparable to the profiles generated from individual cRNA samples, while at the same time, intra- and inter-patient variations were effectively normalized. Their results suggested that pooling or mixing a rather small number of RNA samples from multiple regions of a piece of tissue from an individual and from multiple individuals matched for most variables (age, gender, disease, etc.) can help eliminate variations owing to tissue heterogeneity and provide stringent and robust gene expression data.

Laser capture microdissection (LCM) is also used to help improve sampling[138-141]. LCM is a process by which individual cells can be dissected from a tissue specimen. A typical LCM procedure is as follows (Please refer to Emmert-Buck et al. [138] for a graphical illustration.). A thin, transparent film is applied to the surface of a microscopic slide, which holds a piece of tissue. The film is a thermoplastic film made of ethylene vinyl acetate polymer. Under the microscope, selected areas of film on top of the cells (or tissues) of interest are activated by laser pulse. The activated film has strong focal adhesion power which allows the selected regions to be procured from the tissue. The film with the procured cells is then removed from the slide and cells/tissues adhered are sent to further treatments in DNA, RNA or enzymatic assays. Multiple regions can be procured on a single film in one procedure. The transferred cells or tissues retain their morphological features which can be verified under microscope.

LCM offers an efficient means to isolate cells of interest from other cell types in a tissue. It has been applied successfully to collect samples for DNA microarray based gene expression analysis[139, 141, 142]. One issue associated with LCM is the limited amount of RNA from collected cells. But the problem has been mitigated by amplifying RNA with the T7 RNA polymerase based *in vitro* transcription method (Section 1.2.3.2)[63, 143]. Gene expression profiles from amplified cRNA from LCM have been validated to have high fidelity and reproducibility[139-141]. One potential drawback of LCM is that it uses morphological characteristics to identify cells of interest, and cells with similar morphological features may not have same gene expression profiles. For example, epithelial cells in the adjacent normal prostate tissues resemble normal epithelial cells but genetically their expression profiles share many characteristics with cancerous epithelial cells. This discrepancy may potentially introduce biases into gene expression profiles generated from LCM samples.

1.5.3.2. Tissue handling

Good tissue handling guarantees the quality of the initial RNAs extracted from the tissues and therefore provides an optimum start for target preparation (labeling and/or amplification) as mentioned in Section 1.2.2. For gene expression analysis, tissue handling is of special importance because RNA samples are very sensitive to degradation. Therefore, fresh tissues need to be handled and stored properly to preserve the transcriptomes intact.

There are a couple of major problems in tissue handling which can potentially affect the quality of RNA samples. First, the length of processing time from surgical removal of tissues to collect samples for research use will affect gene expression greatly by inducing ischemia. Ischemia is caused by lack of blood flow to a tissue/organ. Without blood flow, the tissue/organ will be depleted of oxygen and therefore become hypoxic. In the state of ischemia, stress-specific response genes will be transcribed to protect tissues from damage. Prolonged processing time may cause tissues to be ischemic and therefore affect gene expression.

Dash et al.[144] reported the effect of warm ischemia on differential gene expression of radical prostatectomy specimens. In this study, gene expression profiles of tissue samples collected at different time after radical prostatectomy were compared. A number of genes were identified with significant increase of expression at 1 hour after radical prostatectomy. Many of these genes have shown increased expression secondary to ischemic stress, hypoxia. Therefore, after surgical removal or biopsy, tissues should be collected as soon as possible and processed then stored properly if not used immediately.

The means by which tissues are processed after surgical removal or biopsy is also an important factor affecting RNA quality. Typically, tissues are fixed in formalin and embedded into paraffin blocks, which can be preserved for many years and archived. However, RNAs will be degraded by forming cross-links with formalin[135, 136]. Alternatively, tissues can be snap-frozen in liquid nitrogen to avoid RNA degradation. Tissues should be put into solutions, such as RNAlater and ethanol that inhibit RNase[57, 135].

1.5.4. Variations introduced by RNA labeling methods

Various labeling and target preparation approaches have been used in DNA microarray based gene expression studies (Section 1.2.3). A well known recommendation[65, 70] on labeling approaches for DNA microarray based gene expression analysis is that one should never change/combine labeling strategies within a study as the data from more than one method may not be comparable. This indicates the possibility of profound, not yet fully characterized, differences introduced by labeling and/or amplification approaches. Systematic assessments of the performance of these methods[64, 65, 70, 74] have demonstrated that individual labeling methods can significantly influence the gene expression data and data using different amplification approaches on the same type of array may not be directly comparable[145]. In addition, many have pointed out that labeling and target preparation (including labeling and amplification) is one of the many technical variables that can profoundly influence the compatibility of gene expression data across platforms and between laboratories[119, 120, 130]. This section will summarize the comparison studies conducted on various labeling and/or amplification approaches and their effects on gene expression data from DNA microarray based experiments.

Most of the comparison studies use "controlled" studies/experiments where other experiment variables such as the source of total RNA and the type of DNA microarray are controlled to study the effects of different labeling and/or amplification methods on gene expression analysis. Gene expression data from total RNA /mRNA labeled using reverse transcription approach is usually used as the standard for the comparison studies. However, the reverse transcription labeling approach is not necessarily the best method to preserve the initial RNA abundance and complexity. There is significant variation during total RNA extraction or mRNA purification[70, 146]. Therefore, a small number of differentially expressed genes may have to be further verified by Northern blot or quantative RT-PCR. Performance of the labeling approaches is reported as the sensitivity of gene detection and the reproducibility of that detection. Labeling approaches may vary from each other at (1) the gene populations detected as present/differentially expressed in the RNA sample and at (2) the ratio of differential expression, which is expressed as fold change or a statistical metric such as signal-to-noise ratio or t-test value.

A couple of labeling methods, including TSA[72] and the dendrimer approach[76] which cause signal amplification with no target amplification have been recently developed and have not yet been adopted as routine procedure for target labeling in DNA microarray based gene expression analysis. Variations from these methods have not been studied systematically. Stears et al.[76] and Manduchi et al.[74] reported that the dendrimer approach has low background, high signal-to-background ratio, comparable level of reproducibility and ability to detect expression, and requires much less targets for hybridization compared to reverse transcription approach using cDNA microarrays. However, in a comparison study carried out by Richter A et al.[65] using a cDNA microarray comprising genes in iron metabolism regulation, the dendrimer approach failed in detecting any differentially expressed genes. The TSA approach showed high background, moderate accuracy and sensitivity in the same report[65]. In summary, gene expression analysis using these two labeling approaches may be less capable of detecting differentially expressed genes and introduce more biases/variations into gene expression data.

Furthermore, the signal amplification may not recapitulate relative differences in a linear fashion[72].

Many more studies have been conducted on the variations introduced by labeling approaches with target amplification such as the T7 RNA polymerase based in vitro transcription method and SMART PCR method. PCR based methods employ exponential amplification that may distort the initial transcript levels. Therefore are not widely adopted for target preparation in DNA microarray based experiments (also see Section 1.2.3.3 for details)[64, 133].

T7 RNA polymerase based *in vitro* transcription was developed by Gelder and Eberwine[60, 61, 63] and first applied for target preparation in gene expression analysis using Affymetrix GeneChip® arrays by Lockhart et al. in 1996[2]. This method has been adopted with modification as the major target preparation method for various types of DNA microarrays including cDNA arrays and long-and short-oligonucleotide arrays[2, 22, 25, 28, 147-149]. This method amplifies mRNA in a linear fashion and, in theory, introduces little distortion to the initial transcript level in mRNA samples. Therefore, the relative abundance and complexity of mRNA will be maximally preserved in the amplification products. Various studies[2, 61, 69] had proved the linearity of this method. Gelder and colleagues[61] have shown by electrophoresis the distribution of the antisense RNA amplified is similar to the cDNA population from which it was produced. By Northern blot and Southern blot analysis, their results also show that the abundance of the amplified antisense RNA is representative to the parent cDNA. Lockhart and co-authors[2], on the other hands, prove the linearity of amplification with array hybridization results. Poly-A tailed synthetic prokaryotic RNA molecules were spiked into eukaryotic total RNA samples at varies of known concentrations, from low to high, before amplification. The array intensities of each spiked control genes are quantitively related to its concentrations over the entire concentration range[2].

High reproducibility of the T7 based method has been reported in almost every study regardless the type of DNA microarrays used with high correlation coefficients (>0.90) and low average coefficient of variances among replicates (<15%) [25, 64, 69, 73]. Correlations of amplified and non-amplified (total RNA/mRNA) data is in the range of 0.77~0.85[64, 150-152]. In a systematic study conducted by Richter et al.[65] using a custom cDNA array comprising known genes for iron metabolism regulation, the T7 based method yielded the largest number of genes as differentially expressed which is also close to the number of genes presented on the iron chips expected to be expressed in the cell line studied, indicating the T7 based method has the best sensitivity. The reverse transcription labeling method on the other hand detected the smallest number of genes as differentially expressed and therefore target amplification increases sensitivity. When compared to the gene expression data from unamplified mRNA and total RNA (labeled by reverse transcription), most studies show that the T7 driven amplification method detected 80% ~ 94% of genes identified by unamplified mRNA or total RNA as differentially expressed[64, 148, 152, 153]. Similarly to Ritcher et al.'s study [65], the T7 based amplification method usually identifies more differentially expressed genes than unamplified mRNA or total RNA[70, 152, 153]. Genes uniquely called present or differentially expressed in gene expression results from one type of method were further proved to be true by qRT-PCR verification, suggesting an increase of sensitivity using the T7 amplification method. Two-rounds of amplification[60, 63] is necessary to yield enough material for DNA microarray based gene expression analysis. Studies showed that gene expression data from two-round of amplification was relatively comparable (correlation is about 0.93~0.95) to data from standard T7 based method[133, 134, 145, 154]. However, certain variations are introduced by two-rounds of amplification (discussed bellow).

In summary, within the limits investigated to this point, the T7 RNA polymerase based *in vitro* transcription method (one- or two- rounds) has been proven to amplify mRNA linearly and the gene expression profiles from T7 based method demonstrate acceptable/excellent fidelity and reproducibility as well as improved sensitivity. Therefore some scientists suggest the routine usage of RNA amplification for all array based gene expression profiling experiments.[152] However others suggest to use RNA amplification only when the starting material is limited[64] and to verify microarray gene expression results with other independent methods like Northern Blot and qRT-PCR.

For the T7 RNA polymerase based linear *in vitro* transcription, selective sequence amplification could occur even in the case of linear amplification. Sequence specific efficiency of the T7 RNA polymerase may also introduce biases to the initial transcript level at *in vitro* transcription owing to the early termination of transcription because of a very long poly(A) tail or strong secondary structure in sequences[64].

Different variations introduced by the T7 RNA polymerase based amplification approach have been reported in a number of studies[69, 70, 133, 134, 145, 150, 154, 155]. Using an ANOVA model and multiple hypothesis testing, Nygaard et al.[155] reported that target amplification significantly affects differential expression ratio of 10% of the genes studied. Variations introduced by either sequence-dependent or sequence independent manner[69, 70]. Baugh LR et al.[69] reported a reduction of present calls in gene expression profiles of the amplified cRNA by the T7 method because of a high molecular weight product existing in the cRNA products. This product is produced by the T7 RNA polymerase in the presence of the carried-over oligo (dT)-T7 primers. Li Y et al.[94] reported that systematic biases can be introduced by RNA amplification method in a sequence-dependent rather than copy-number dependent fashion. The authors also reported sequence dependent biases and 5' under representation introduced by T7 based RNA amplification.

Variations are also observed in the amplification products where the distribution of cRNA products changed compared to the distribution of total RNA or mRNA, indicating the possible distortion of relative transcript abundance caused by amplification. For example, several reported showed that amplified products have smaller range than purified mRNA and there is a shift toward small transcript size in distribution (we also observed the shift of transcript size distribution in Chapter IV, Section 4.4.3)[153, 156]. More shifts were observed with two-round of amplification[134, 154]. Polacek et al. speculated that this shift is because of early termination of reverse transcription and/or in vitro transcription of the T7 based amplification method[153]. Spicess et al. on the other hand, thought this is owing to the cRNA degradation by T7 RNA polymerase and suggested that one should never compare gene expression results with cRNAs from different amplification times[156]. A number of studies also demonstrated the decreases of intensity values of certain transcripts in gene expression data caused by 5' truncation at the T7 target amplification process[69, 133, 145, 148]. For such a transcript, the 3'/5' intensity ratio, if there are both probes designed from 5'- and 3'- end sequences, will subsequently increased. This is quite a prominent variation observed in gene expression results from two-round amplification[133, 145]. In addition, Gold et al. speculated the 5' truncation in amplification is the cause of the observed decrease of sensitivity after two-round of amplification[134].

Efforts had been made to correct some of the variations aforementioned and to improve gene expression data from amplification. For example, a template switch strategy has been used to synthesize of second strand cDNA to ensure the generation of full-length ds-cDNA[148, 151]. However, Zhao et al.[151] demonstrated that template switch did not help to improve fidelity of gene expression data with target amplification. Baugh et al.[69] improved gene expression results from amplified targets by removing the template-independent, high molecular weight products which is caused by the excess amount of oligo-(dT)-T7 primer during amplification. However, variation can not be 100% removed from amplified products indicating there are uncharacterized sources responsible for the observed variations other than 5'-end truncation and template-independent high molecular weight products. Understanding the possible sources of variations and the impacts of the variations to gene expression results will also help on better study design and cross-platform comparison.

In addition, there are many "versions" of the T7 RNA polymerase based *in vitro* amplification methods available, each of which is designed to produce optimized array hybridization results as claimed by its manufacture. It is intuitive to speculate that these methods may introduce variations into gene expression results even though they are all derived from the original T7 based method. However, up to now, there is no systematic comparison of different methods.

Chapter IV in this thesis described a comparison study of three popularly used T7 based methods. Results from this comparison study demonstrated the existence of significant variations introduced across different T7 based target amplification methods. In addition, results also demonstrated that both the number of biotinylated nucleotides used for labeling and the reaction time of *in vitro* transcription are responsible for the observed variations.

2. CHAPTER II Effects of analysis methods on the supervised classification of prostate tissue samples using microarray data: implications of variation introduced by sampling and tissue handling

Resubmitted to BMC Bioinformatics

2.1. Abstract

Tumor classification and class prediction has become one important application of microarray technology. In this study, we examine the classification of prostate tumor tissue and normal (non-tumor) prostate tissue using three different classification methods (Boosted Decision Tree based on the C4.5 algorithm, Support Vector Machines and Weighted Gene Voting) at various levels of feature selection. In addition, we were able to divide the normal (non-tumor) samples into different types (normal prostate tissue from cancer prostatectomies and normal prostate tissue from tumor free organ donors), and examine the effect of using one or the other in the analysis. Our results indicate the boosted decision tree results were as good if not better as the classification produced by more accepted microarray classification methods such as support vector machines and weighted gene voting. Significantly, the type of 'normal tissue' used in the analysis had a significant impact on the accuracy of the classifiers, indicating that the sample selection and tissue processing may be much more important than the specific analysis method used in the interpretation of microarray data.

2.2.Introduction

Over the past several years, gene expression microarray technology has been successfully applied in prostate research. Numerous groups have reported the discovery of individual genes associated with prostate cancer[38, 45, 139, 157, 158], the reliable discrimination of tumor and benign samples[38], and the correlation of gene expression with traditional morphologic metrics of tumor progression such as Gleason grade[38]. In addition, the possible discovery of prostate cancer sub-types with varying degrees of aggressiveness has been reported[40].

One of the most critical microarray applications is tumor classification and class prediction – the task of classifying samples into known diagnostic classes. Compared to unsupervised methods such as clustering[98], supervised classification methods are preferred for performing these tasks because such methods takes advantage of existing information and domain knowledge and therefore should create a more accurate (and reliable) classifier. Supervised classification methods take training sets in which the expression of each gene and diagnostic class (tumor, non-tumor, etc) of each sample is known and use that information to build a classifier that can predict the diagnostic class of new tissue samples based on their gene expression data.

Supervised classification will be particularly important in the pathology laboratory, when they are faced with the task of classifying a clinical specimen into one or more diagnostic categories (for example, benign, pre-malignant, and malignant) based on microarray results. It will also be important for predicting progression and patient survival. This scenario is becoming more and more realistic, as an increasing number of hospitals (including the University of Pittsburgh Medical Center) are implementing clinical laboratories to use microarray technology for diagnostic, predictive and prognostic classification (on an experimental basis).

There are however, a number of open questions surrounding the supervised classification of tissue samples in general and prostate samples in particular. For example:

(1) There are multiple supervised classification methods in use including the Nearest Neighbor classifiers[38], the Weighted Gene Voting method[32], Support Vector Machines[112, 113], Shrunken Centroids[114], Fishers Linear Discriminant Analysis (FLDA)[111], and Decision Tree Based Methods using Recursive Partitioning[115]. There is limited information on which are most accurate and appropriate for prostate tissue classification.

(2) Most microarray experiments measure a large number of genes (more than 10,000) on a few dozen tissue samples. This high dimensionality makes analysis and interpretation difficult. The expression of the great majority of genes (reported by a microarray) does not change between specimen classes. These genes therefore do not contribute a differential "signal" but do contribute "noise", making the classification process less effective and potentially more costly. For these reasons, there has been interest in "feature selection" as a pre-processing step before the application of classification algorithms. Numerous statistical methods[32, 109] have been developed to reduce the dimensionality of microarray data by selecting only genes that are significantly expressed between specimen classes. However, the effect of feature selection should be examined systematically.

(3) The microarray experiment itself is difficult to perform and control. Numerous experimental factors such as sampling, tissue handling[144] and storage conditions can affect results. The adult human prostate is an architecturally complex, hormonally sensitive organ that continues to evolve over a person's lifetime. Significantly, different

topological zones in the prostate manifest different biology - adenocarcinoma, a common finding in the organ's periphery seldom involves the central or transitional zones. Most intriguing, published studies using a variety of techniques as diverse as chromosomal analysis[159], SAGE[160], ploidy analysis[161, 162], and image analysis[163-165] have shown molecular and morphologic abnormalities in normal appearing prostate adjacent to adenocarcinoma.

In this study, we evaluate the classification and prediction of prostate tissue samples from three independent data sets. We used three classification methods (Boosted Decision Tree (BDT)[166], Support Vector Machines (SVMs)[112, 113], and Weighed Gene Voting (WGV)[32]). We examined the results at different levels of feature reduction and, most importantly, compared tumor samples against different types of "normal" baselines. Our results indicate that the classification methods performed in a fairly similar manner (both well or poorly depending on the samples compared), feature selection can have an important impact on accuracy of BDT and, in these prostate data sets, tissue sampling and processing methods may be much more important than the specific statistical methods used in the analysis process.

2.3.Materials and methods

2.3.1. Data sets

Three data sets were used in this study. The data sets had in common the use of tumorous and normal appearing human prostate tissue. However, the data sets were generated at different institutions using different populations and different sampling techniques. These differences were particularly significant in the selection of normal tissue. While tumor samples were taken from prostate cancer patients, normal samples were taken from both the normal appearing regions of prostatectomy specimens (Patient Normal), and from normal appearing areas of tumor free prostates taken from organ donors (Donor Normal). Table 2.1 gives the clinical and pathologic features of patients, donors, and specimens.

The Primary Data set

The primary dataset (Table 2.1) used in this study was generated by scientists at University of Pittsburgh Medical Center (UPMC)[40]. There were three sample types used in this data set. Eighteen "Donor Normal" samples were taken from prostates harvested from eighteen organ donors. These prostates were certified to be free of tumor by UPMC pathologists and the samples were felt to be histologically "normal" (no prostatic hyperplasia, inflammation, etc.) Samples were also taken from patients undergoing radical prostatectomy for prostate adenocarcinoma. From sixty-three patients, 60 tumor samples and 63 samples of normal appearing prostate adjacent to tumor (Patient Normal) were analyzed. In all, 141 samples (18 Donor Normals, 60 Tumors and 63 Patient Normals) were run on the Affymetrix U95Av2 chip. The raw expression data was then analyzed using MAS5.0 software (Affymetrix Inc.) and normalized using a global normalization approach. This data is referred to as the "PITT" data set in the following discussions. Please also refer to **Appendix A** for a related study on this data set.

The Two Independent Test Data sets

Two previously published prostate tumor datasets (Table 2.1), referred as "Singh" data set[38] and "Welsh" data set[39], were obtained from public domain or given kindly by the authors respectively. The Singh data set was generated from patient prostatectomy specimens. There were 52 histologic tumor samples and 50 samples taken from normal appearing areas of the prostatectomy adenocarcinoma (Patient Normals) using Affymetrix U95Av2 chip. The Welsh data set was generated by using Affymetrix U95A chip (12600 probe sets are overlapped between U95A and U95Av2 chips) from 24 histologic tumors and 8 normal tissue samples

(Patient Normals). Normal samples were obtained from 8 of the specimens with tumor. The raw expression data (".dat" files and ".cel" files) of these two datasets were reanalyzed using MAS5.0 software and normalized by the same global normalization method as that used for PITT data set. This procedure makes the three data sets comparable for the subsequent analysis.

Table 2.1 Clinical and pathological features of the prostate tissue samples in the three data sets.

		PITT		Singh		Welsh	
Tumor Samples							
No. of Tumor Samples		60		52		25 (24 unique samples)	
Features of Tumor Samples							
Pathological Stage	**T2a**	2	3.3%	7	13.5%	5	20.0%
	T2b	21	35%	25	48.1%	6	24.0%
	T3	0		0		1	4.0%
	T3a	24	40.0%	16	30.8%	6	24.0%
	T3b	11	18.3%	4	7.7%	2	8.0%
	T4	1	1.7%	0	0.0%	0	
	T4a	1	1.7%	0	0.0%	0	
	Tx	0		0		4	16.0%
Gleason Grade	**5**	2	3.3%	4	7.7%	1	4.0%
	6	13	21.7%	15	28.8%	7	28.0%
	7	28	46.7%	29	55.8%	9	36.0%
	8	5	8.3%	2	3.8%	5	20.0%
	9	12	20.0%	2	3.8%	2	8.0%
Age	**40-49**	4	6.7%	3	5.8%	2	8.0%
	50-59	19	31.7%	24	46.2%	8	32.0%
	60-69	24	40.0%	22	42.3%	14	56.0%
	70-79	13	21.7%	3	5.8%	0	
Other information	PSA values before and after prostatectomy; recurrent; vital status; seminal vesicle invasion; extension through capsule; etc.			PSA value; Gland volume; extension through capsule; positive surgical margin; seminal vesicle invasion; recurrent; non-recurrent; etc.		Selected transcript levels (PSA, Hepsin, MIC-1); percentage of various cell types from the section adjacent to the tissue profiled; etc.	
Non-Tumor Samples							
No. of Non-Tumor Samples		81		50		8	
Donor		18		0		0	
Features of Donor samples							
Age	**<10**	1	5.6%				
	10-19	4	22.2%				
	20-29	5	27.8%				
	30-39	1	5.6%				
	40-49	3	16.7%				
	50-59	4	22.2%				
Patient Normal		63		50		8	
Features of NAT samples							
Age	**40-49**	4	6.3%	3	6.0%	0	
	50-59	20	31.7%	21	42.0%	4	50.0%
	60-69	26	41.3%	23	46.0%	4	50.0%
	70-79	13	20.6%	3	6.0%	0	

2.3.2. Classification approaches

Three classification methods were used in this study, the Boosted Decision Tree approach (BDT)[167], Support Vector Machines (SVMs)[112, 113] and the Weighted Gene Voting (WGV) method[32]. The goal was to introduce the decision tree learning method based on C4.5 algorithm and compare its performance with the other two commonly used, successfully applied, classification methods.

C4.5 Decision tree learning (DT)

The decision tree learning programs used in this study were C4.5 Release 8 (Quinlan, 1993), which is a freeware, and its related commercial product C5.0/See5 Release 1.17 (RuleQuest Research Pty Ltd., Australia).

C4.5 uses the “divide and conquer” technique (also called recursive partitioning) to construct a decision tree from the training set. Each member of the training set belongs to a single class (for example benign or malignant). Each member also has one or more attribute-value vectors where the values are mutually exclusive (for example, invasive or not invasive). For a continuous attribute such as the expression level of a gene or the serum level of a marker, cut offs are uses to construct mutually exclusive outcomes (e.g. > 5.0 and <= 5.0). To create a tree, an attribute is selected based on the “information gain” and the training set is split (partitioned) into subsets where all members of a subset have the same value for the given attribute. This process is applied recursively until all subsets contain members of a single class. A simplified set of rules defining each leaf (class) can further be derived after the tree is grown. These rules are not expressed as a tree structure and may be more attractive to the user [168]. Information gain [168] measures of how well a given attribute/gene separates the training set into subsets. It is defined as the expected reduction in entropy caused by partitioning the examples according to this gene.

$$\text{Gain } (S, A) \equiv \text{Entropy } (S) - \sum |Sv| \text{ Entropy}(Sv) / |S|;$$

$$v \in \text{Values } (A) \qquad (1)$$

Where A is a gene, Values (A) is the set of all possible values for attribute/gene A, and Sv is the subset of S for which attribute/gene A has value v. Entropy [168] in equation (1) represents the purity of samples in a subset and is defined as:

$$\text{Entropy } (S) \equiv \sum i -p_i \log2 p_i \quad i = 1, \ldots, c \quad (2)$$

The decision tree built from training examples will be used to predict the classification of new examples (in the "testing" set). When a new example comes into the tree, it starts to traverse the decision tree from the root node. At each node, the corresponding test is performed on it. The outcome of the tested attribute determines the branch on which the sample will descend to the next test/node. This process ends when a leaf is encountered and the classification of this leaf determines the predicted class of this example.

The boosting approach

The See5/C5.0 Release 1.17 (RuleQuest Research Pty Ltd., Australia) has implemented adaptive boosting based on the work of Rob Schapire and Yoav Freund[169]. Boosting is a general method in machine learning to improve the accuracy of any given learning algorithm (not just decision trees) by generating and combining multiple inaccurate classifiers/rules. Each training example is given a weight. In each trial, the boosting algorithm generates a classifier based on all the weighted training examples and assigns a new weight to each of them based on the classification results such that misclassified examples get more weight as the classification proceeds. Adjusting weight at each round forces the learning algorithm to focus on the different examples and thus generates different classifiers. At the end, all the classifiers are combined by voting according to their accuracies to create a compound classifier[169]. In this study, the number of trials used was increased from 0 to 50 with a pace of 5 trials each time.

Support vector machines (SVMs)

Gist software tool for support vector machine classification version 2.0.2 is used in this study. It is developed by William Stafford Nobel in the Department of Computer Science at Columbia University and by Paul Pavlidis in the Columbia Genome Center (http://microarray.cpmc.columbia.edu/gist/).

An SVM attempts to computes a multi-dimensional plane (a hyper-plane) that separates all members of two different classes in the multidimensional microarray data set. Specifically the SVM computes a “maximal margin hyper-plane” that separates the two classes in the training set such that each training example has a maximum distance between it and the hyper-plane. When training examples are not linearly separable in the input space, i.e. SVM can’t find a maximal margin hyper-plane that can completely separate the classes in the input space, SVM uses a kernel function to map all the training examples to a higher-dimensional feature space where a maximal margin hyper-plane can be located and thus training examples can be linearly separated by it[112, 113]. In this study, only the simple dot-product kernel is used. In addition, soft-margin has been applied by tuning the diagonal factor (DF) to control the training error[113]. The range of DF was set to 0, 0.01, 0.1, 0.5, 1, 2, 5 and 10.

Weighted gene voting (WGV)

Weighted Gene Voting (WGV) is a method for binary classification first proposed by Golub et al. (1999)[32]. In WGV, genes are ranked based on a selection metric such as the signal-to-noise metric[32]. Then for the selected genes, each gene casts a vote for class 1 or 2.

Finally, a class vote for each example calculates as the summation of all the votes from the selected genes.

2.3.3. Feature selection

For our feature selection criterion, we use the signal-to-noise (S2N) metric developed by Golub et al. (1999)[32]. Given “n genes” and “m tissue samples”, each of the m tissue samples is labeled either 1or 2 depending on the comparison at hand. (e.g. Tumor (1) v Donor Normal (2)). For each of the n genes, we calculate the mean, μ1 and μ2, and standard deviation, σ1 and σ2, for the samples labeled 1 or 2 respectively. The signal-to-noise metric is calculated as:

$S2N_{ij} = (\mu_{ij}-\mu_{ij}) / (\sigma_{ij}+\sigma_{ij})$; i = 1, …, n, and j = 1, 2;

The S2N metric gives the highest and lowest scores for genes whose expression levels differ most on average in the two classes while also favoring those with small deviation in scores in respective classes. Genes are then ranked by S2N. The most positive and most negative genes are genes selected as the most differentially expressed, and therefore the top ranked, features used in this study.

2.3.4. Experimental method

Purposes of this study were to compare the effectiveness of the three classification methods (BDT, SVM, and WGV) across a series of data sets and to explore the classification tasks with gene expression data generated from human cancer specimens. The Experimental method can be described as follows (figure 2.1).

The initial phase of the study (Figure2 1a) concentrated on the binary classification of Tumor and Non-Tumor samples in the PITT data set. As we described before, Non-Tumor samples in the PITT data set consist of Patient Normal samples and Donor Normal Samples. We applied each of the three classification methods (BDT, SVM, and WGV) to this comparison using a “leave-one-out” cross-validation (LOOCV) format (Figure2 1b). Initially, this was done against the entire feature set (~ 12,600 feature/sample).

Returning to the initial data set, in the leave-one-out format, we used the S2N method (vide supra) to rank each feature (gene) for differential expression in the comparison of Tumor v Non-Tumor samples (i.e. gene selection is also subjected to cross-validation). We then progressively re-applied the classification methods (in the same leave-one-out format) to data sets limited to the most differentially expressed features (genes) in the comparison (as determined by the S2N algorithm). This was done at various levels of stringency, resulting in data sets that represented the 2000, 1000, 500, 200, 100, 50, 20 and 10 most differentially expressed features in the comparison. At each level of stringency, equal numbers of the most up and down regulated genes were selected to compose the pre-selection features (e.g. the 2000 most significantly expressed genes contain 1000 most up regulated genes and 1000 most down regulated genes). The results of each classifier, at each level of pre-selection of features, were then compared (figure 2.1b).

In the second phase of the analysis (figure 2.1a), Patient Normal samples and Donor Normal samples from the Non-Tumor category in the PITT data set were used separately to build classifiers with Tumor samples (in the PITT data set) separately. This gave us two types of comparisons: Tumor v Patient Normal and Tumor v Donor Normal. As in the phase 1, the three classification methods were first used to generate classifiers using the full feature (gene) set and a series of smaller feature (gene) sets selected for differential expression using the S2N metric in a LOOCV fashion. The Tumor v Patient Normal and Tumor v Donor Normal) classifiers were further investigated by using the PITT data set as a training set and two, previously published, independent data sets as test data sets. As in phase 1, the three methods were first used to generate and test classifiers using the full feature sets and a series of smaller feature sets selected for differential expression using the S2N algorithm. It is important to note that this time the most significantly expressed genes were selected using all training samples in the comparison because the performance of the classifiers was assessed by its ability to predict classes of independent samples (Figure 2.1c).

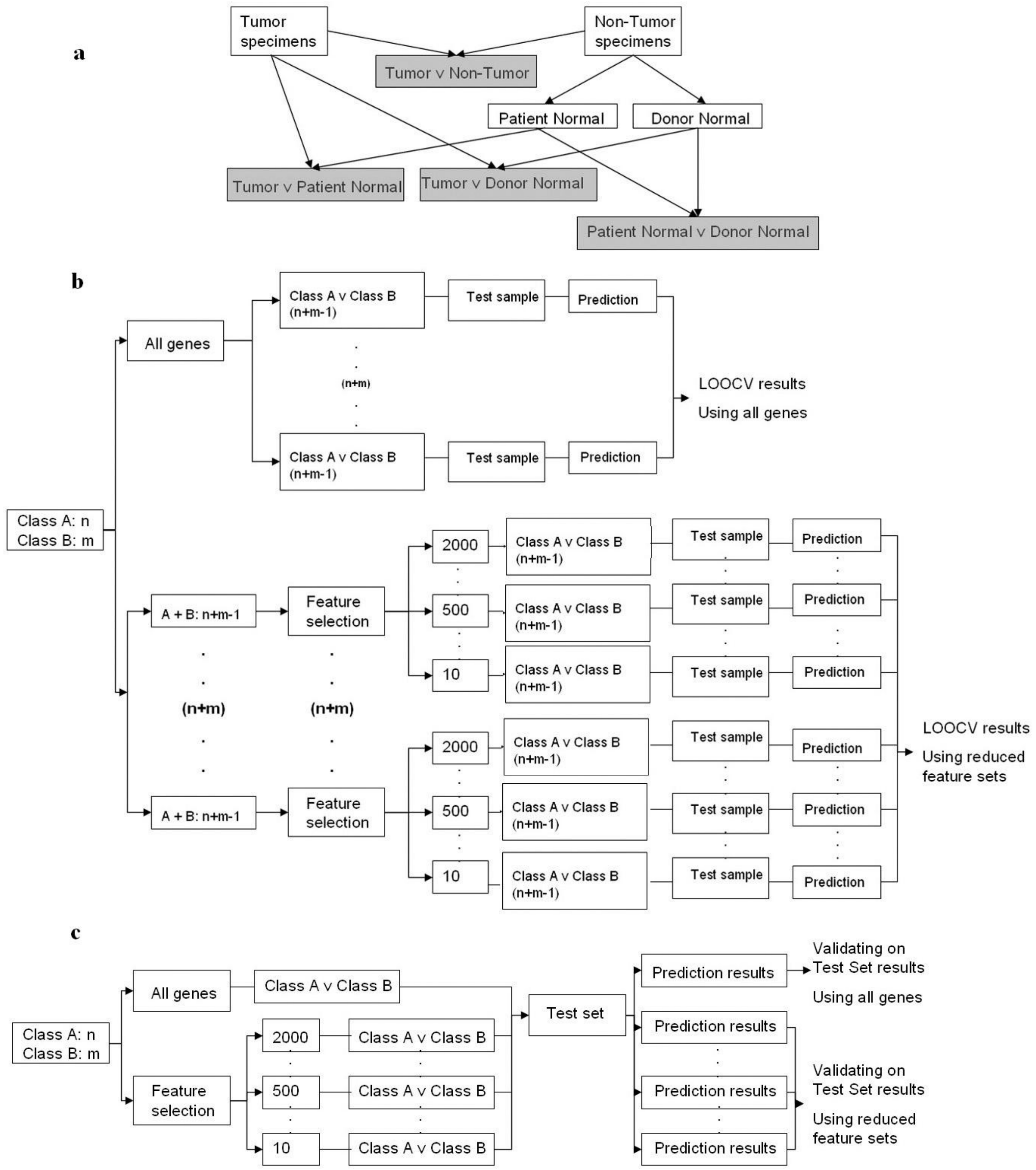


Figure 2.1 Depicts the experimental method in detail.

Figure 2.1 legend: a) Illustrates the three phases of the experimental design. Four types of classifiers were built from three types prostate tissue samples in PITT data set. b) shows the procedures for the leave-one-out cross-validation (LOOCV) experiments when combining with feature selection. Two classes of tissue samples, class A and class B, are used to construct classifiers where class A has n tissue samples and class B has m tissue sample. When performing LOOCV using the full set of genes (all genes), a classifier (Class A v Class B) was built with (n+m-1) tissue samples and then used to predict the class membership of the tissue sample left out. This process was repeated (n+m) times until all tissue samples had been left out once. When feature selection is performed in combination with LOOCV, feature selection is performed using the (n+m-1) tissue samples to generate a series reduced feature sets consisting of a pre-defined number (2000, 1000, 500, 200, 100, 50, 20, 10) of most significantly expressed genes (i.e. feature selection is also subject to leave-one-out cross validation). Classifiers (Class A v Class B) are constructed use the reduced feature sets and used to predict the classification of the one left-out tissue sample. This process is performed (n+m) times as each sample needs to be left out once. Prediction results for each left-out tissue sample are summarized to give the LOOCV results using the reduced feature sets. LOOCV results are shown in Figure 2.2c and Table 2.2) Shows the procedure of validating classifiers on independent data. The classifiers (Class A v Class B) are built from both the full feature set (all genes) and the reduced feature sets generated by feature selection procedure and tested on predicting the class membership of the tissue samples in the same test data set. Prediction results are summarized in Figure 2.2 and the best results are listed in Table 2.3.

2.3.5. Performance evaluation

Four types of classifiers were generated for binary classification of the prostate specimens: a classifier for Tumor v Non-Tumor samples, a classifier for Tumor v Donor Normal samples, a classifier for Tumor v Patient Normal samples and a classifier for Patient Normal v Donor Normal Samples. Tumor samples are always considered as "positive" and the other type of tissue samples, Non-Tumor, Patient Normal, or Donor Normal, are "negative" samples in these comparisons. In the Patient Normal v Donor Normal comparison, Patient Normal is "positive" and Donor Normal is "negative". The "positive" and "negative" assignments are arbitrary but necessary for measuring the performance of a classifier discussed in the following paragraph.

The performance of each classifier was measured by examining how well it identified tissue samples belonging to its two classes in the training set itself while doing LOOCV or in the test sets since the class assignment of each tissue samples were known prior of learning. Using the classifier for Tumor v Non-Tumor as an example, each tissue sample can be categorized in one of four ways: the true "positives" (TP) and the true "negatives" (TN) are the Tumor and the Non-Tumor tissue samples, respectively, according to both the classifier and their true class assignments; the false "negatives" (FN) are Tumor tissue samples classified as Non-Tumor tissue according to the classifier; the false "positives" (FP) are Non-Tumor tissue samples classified as Tumor by the classifier. The number of samples in each category is reported for each of binary classifier. The overall performance is measured by the "Accuracy" of classification. Accuracy is calculated as Accuracy = (TP + TN) /Sum where Sum = (TP + TN + FP + FN).

2.4.Results

2.4.1. Tumor v Non-Tumor classification

Table 2.2 shows the best leave-one-out cross-validation (LOOCV) results of the classifiers built from Tumor and Non-Tumor tissue samples in the PITT dataset using the three classification methods – BDT, SVMs and WGV. Feature reduction was also applied and cross validated. (LOOCV results at different levels of feature selection are given in Figure 2.2a.) Parameters of each method were tuned to get the best classification results. All classifiers gave comparable accuracies, between 0.74 and 0.8, while using reduced feature sets. The best LOOCV result was generated by boosted decision tree (BDT) method where 41 of the 60 Tumor samples and 72 of the 81 Non-Tumor samples were classified correctly.

These results illustrate that Tumor tissue samples cannot be classified accurately in LOOCV when Non-Tumor samples were used as the "baseline" to build classifiers (since the best LOOCV result is 0.8 as discussed above). In addition, the poor performance of the classifiers for Tumor v Non-Tumor makes the validation on independent data sets less essential. We did not further investigate prostate tissue classification with this classifier.

2.4.2. Tumor v Patient Normal classification and Tumor v Donor Normal classification

Non-Tumor tissue samples were separated into two groups – Patient Normal samples and Donor Normal samples based on their tissue origins. Tissue samples in each group were then used to build classifiers against Tumor tissue samples. Thus two types of binary classifiers were generated – a classifier for Tumor v Patient Normal and a classifier for Tumor v Donor Normal.

Table 2.2 LOOCV results of the four types of classifiers built by PITT data set using three methods.

Classifiers	Tumor vs. Non-Tumor			Tumor v Patient Normal			Tumor v Donor Normal			Patient Normal v Donor Normal		
classification methods	BDT	SVM	WGV	BDT	SVM	WGV	BDT	SVM	WGV	BDT	SVM	WGV
Parameters	Boost 0	DF 0.1	NA	Boost 30	DF 0.5	NA	Boost 0	DF 0.1	NA	Boost 5	DF 0.1	NA
Feature	20	50	100	20	200	1000	20*	50	100	500	200	100
FP	9	21	19	13	21	18	1	1	1	4	0	1
FN	19	15	16	16	16	17	3	0	1	1	2	4
TP	41	45	44	44	44	43	57	60	59	62	61	59
TN	72	60	62	50	42	45	17	17	17	14	18	17
AC	0.80	0.74	0.75	0.76	0.70	0.72	0.95	0.99	0.97	0.94	0.98	0.94

BDT: boosted decision tree; SVM: support vector machine; WGV: weighted gene voting; FP: false positive; FN: false negative; TP: true positive; TN: true negative; AC: accuracy; DF: diagonal factor.

*: When the classification results are the same at different levels of feature selection, the smallest number of genes is reported.

2.4.2.1. LOOCV results

LOOCV results at each level of feature selection are given in Figure 2.2b and Figure 2.2c whereas the best results generated from each method are shown in Table 2.2.

The classifiers for Tumor v Patient Normal did not perform well in LOOCV experiments. All three methods gave similarly low accuracies (< 0.8). The BDT classifier gave the highest accuracy of 0.76 when applied to the 20 most significantly expressed genes. 16 of the 60 Tumor tissue samples and 13 of the 63 Patient Normal tissue samples were misclassified as Patient Normal and Tumor respectively. SVMs gave the lowest accuracy of 0.7 (using 200 most significantly expressed genes. In aggregate, Tumor v Patient Normal classifier results were no better and actually even a little worse than the LOOCV results from the classifiers for Tumor v Non-Tumor.

The classifier for Tumor v Donor Normal, however, yield significantly high accuracies in LOOCV study. All three methods gave accuracies above 0.9. Classifiers built by BDT had the lowest accuracy of 0.95 and the classifiers by SVMs gave the highest accuracy of 0.99 where only 1 Donor Normal tissue sample (and no Tumor samples) was misclassified.

These results show that Tumor samples can be classified correctly in LOOCV when Donor Normal samples were used as "baseline" to create the classifiers. Patient Normal samples on the other hand were difficult to separate from Tumor samples regardless of classification method or degree of feature selection.

2.4.2.2. Validation on independent data

The two types of classifiers (Tumor v Donor Normal and Tumor v Patient Normal) built from the three methods were further validated on two independent data sets – the Singh data set and the Welsh data set. Although the classifiers for Tumor v Patient Normal had demonstrated poor prediction accuracy in LOOCV, we still attempted to validate the Tumor v Patient Normal classifiers on the independent data sets. The results were compared with those from the classifiers built on the Tumor v Donor Normal comparison (Table 2.3, Figure 2.2e, f, h and i).

As we expected, the classifiers for Tumor v Patient Normal (Table 2.3 and Figure 2.2e) generated from all three methods did not give good accuracy on predicting the class membership of the Tumor and Patient Normal tissue samples in the Singh data set. The lowest accuracy is 0.73 and is obtained by the classifier built by WGV. The highest accuracy was 0.84 by the classifiers built by SVMs where 11 of the 52 tumor tissue samples and 5 of the 50 Patient Normal tissue samples in the Singh data set were misclassified as Patient Normal and Tumor correspondingly.

Significantly, in the original paper by Singh et al.[38], Tumor samples and Patient Normal samples in that data set had been reported to show significant differences in gene expression and were classified successfully (LOOCV accuracy was 0.95) by a nearest neighbor classifier.

However, such two notably distinct types of tissue samples were inseparable by our classifiers built using the Tumor and Patient Normal tissue samples in the PITT data set regardless of the classification method employed.

Table 2.3 Validation results of three types of classifiers built by PITT data set built by three methods on two independent data sets.

Classifiers	Tumor v Patient Normal			Tumor v Donor Normal			Patient Normal v Donor Normal		
Validation on Singh Data Set									
Classification methods	**BDT**	**SVM**	**WGV**	**BDT**	**SVM**	**WGV**	**BDT**	**SVM**	**WGV**
Parameters	Boost 35	DF 1	NA	Boost 10	DF 0.01	NA	Boost 5	DF 0	NA
Feature	200	200	20	10	20	100	10	20	500
FP	10	5	24	0	0	0	0	0	0
FN	13	11	4	0	1	0	2	4	0
TP	39	41	48	52	51	52	48	46	50
TN	40	45	26	0	0	0	0	0	0
AC	0.77	0.84	0.73	1	0.98	1	0.96	0.92	1
Validation on Welsh Data Set									
Classification methods	**BDT**	**SVM**	**WGV**	**BDT**	**SVM**	**WGV**	**BDT**	**SVM**	**WGV**
Parameters	Boost 15	DF 5	NA	Boost 5	DF 0	NA	Boost 0	DF 0.1	NA
Feature	1000*	50	100	10	10	50	10	10	100
FP	0	0	0	0	0	0	0	0	0
FN	1	1	1	0	0	0	0	0	0
TP	24	24	24	25	25	25	8	8	8
TN	8	8	8	0	0	0	0	0	0
AC	0.97	0.97	0.97	1	1	1	1	1	1

BDT: boosted decision tree; SVM: support vector machine; WGV: weighted gene voting; FP: false positive; FN: false negative; TP: true positive; TN: true negative; AC: accuracy; DF: diagonal factor.

*: When the classification results are the same at different levels of feature selection, the smallest number of genes is reported

On the other hand, the validation of the Tumor v Patient Normal classifiers on Welsh data set (Table 2.3 and Figure 2.2h) was very successful. Classifiers from all three methods gave the same high accuracy of 0.97. Only one Tumor tissue sample from the Welsh data set was misclassified and all Patient Normal tissue samples were classified perfectly as Patient Normal. This result is contradictory to the poor prediction results on Singh data set discussed above and was somewhat surprising.

Welsh data set was also used by authors of the Singh data set to validate a nearest neighbor classifier for Tumor v Patient Normal tissue samples in the Singh data set. The prediction accuracy was also notable high (Accuracy was 86% using 16-gene model despite a 10-fold difference in overall microarray intensity between these data sets.)[38]. However similar classifier developed using the Singh data set failed to predict the classes of Tumor and Patient Normal tissue samples in the PITT data set accurately (accuracy ~ 72%, data not shown).

The finding that classifiers built on the PITT Tumor v Patient Normal and classifiers built on the Singh Tumor v Patient Normal both validate in the Welsh data set yet do not validate against each other's data set suggests that there maybe a limitation in use of the Welsh data set for validation. The Welsh data set has good internal control and limited noise, but is rather small, containing 25 Tumor tissue samples and 8 Patient Normal samples and therefore may not represent all the characteristics of the Tumor and Patient Normal prostate tissue population. Thus, the prediction results on this data set may not provide enough information about the accuracy of Tumor v Patient Normal classifiers.

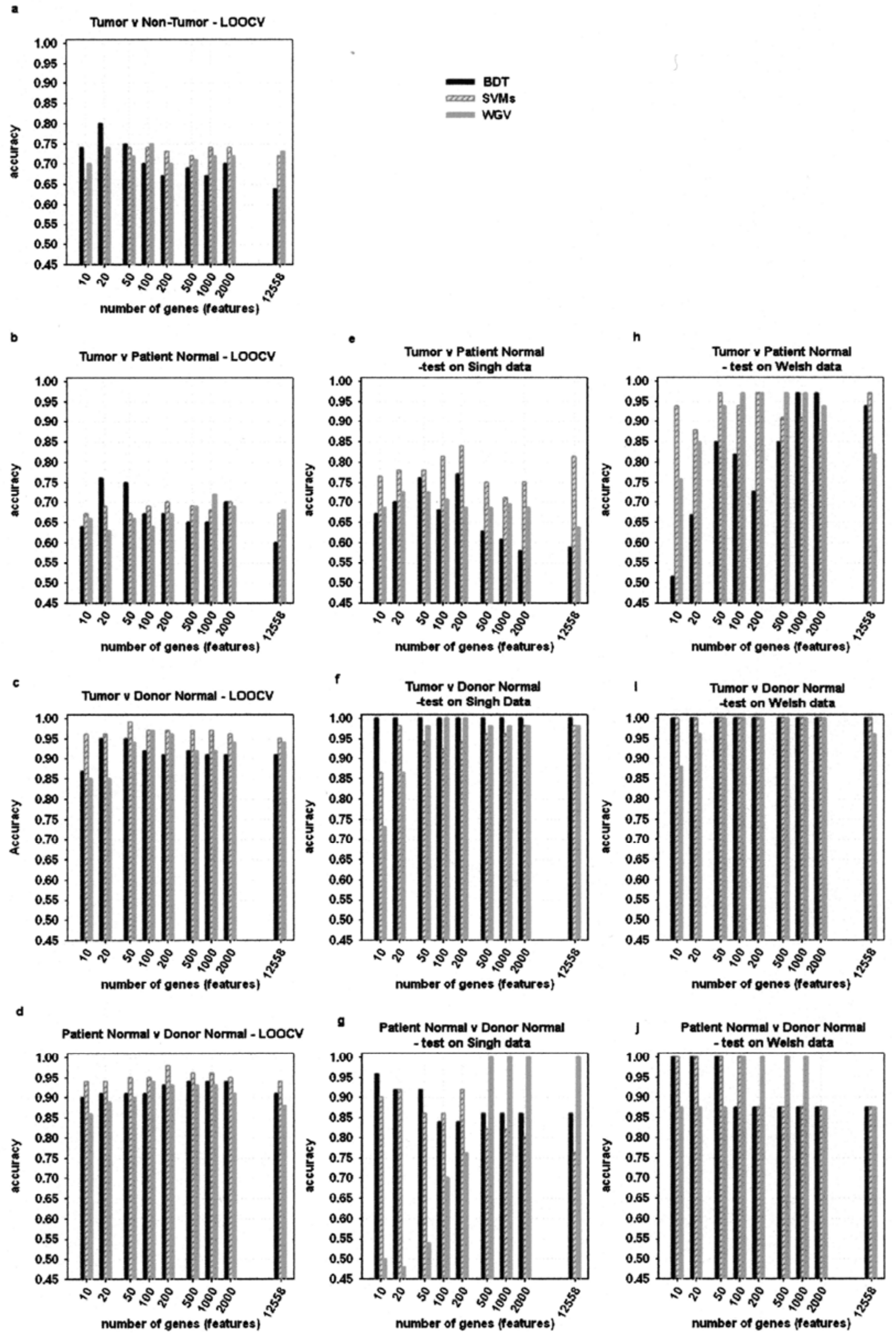


Figure 2.2 Bar charts on the comparisons of accuracies of the classifiers built by the three methods at different level of feature selection.

Figure 2.2 legend: x axis is the number of genes used in common log scale (log 10); y axis represents accuracy. DBT: boosted decision tree, SVMs: support vector machines; WGV: weighted gene voting. Accuracies are represented by bars. The accuracy bars of the same type of classifiers built by the three methods using the same number of significant genes are grouped together.

Since neither the Singh data set nor the Welsh data set have Donor Normal tissue samples, only the Tumor tissue samples were used to validate the performance of the classifier for Tumor v Donor Normal built by the three classification methods from PITT data set. Therefore, in Table 2.3, the true negative and false positive categories are empty and accuracies were calculated based on the classification results of Tumor tissue samples in the two data sets. The lack of Donor Normal tissue samples in the two independent test data sets limits the complete evaluation of the performance of this classifier. In particular, no information will be provided on the specificity of this classifier giving that specificity is measured by dividing true negatives with all "negative" samples. Here, the specificity of this classification is defined as the probability that a tissue sample is not classified as Tumor given that it is not a Tumor tissue sample (it is a Donor Normal tissue sample). However, we know of no other prostate cancer data set in the public domain that has Donor Normal tissue samples. LOOCV results of the classifier for Tumor v Donor built by PITT data set (reported above) can provide nearly unbiased error estimate. Furthermore, the Tumors are more difficult to classify and the classifiers' performance on Tumor tissue samples is more interesting to us. We report the validation results in Table 2.3 and Figure 2.2.

The classifiers for Tumor v Donor Normal built with PITT data were very successful in predicting the classes of the Tumor tissue samples in Singh and Welsh data sets (Table 2.3 and Figure 2.2f, 2i). Both the classifier built by BDT and the classifier built by WGV perfectly classified all Tumor tissue samples while the classifier by SVMs misclassified one Tumor tissue sample in Singh data set. All the classifiers built by the three methods perfectly predicted the classification of all Tumor tissue samples in the Welsh data set. Compared with the validation results of the classifier for Tumor v Patient Normal discussed earlier, these results suggest that

Tumor tissue samples in the two independent data sets are classified perfectly by the classifiers using Donor Normal as “baseline”.

Both the results of LOOCV and results of validation on the two independent data sets imply that Donor Normal is a better “baseline” than Patient Normal. The Tumor and Patient Normal tissue samples in the PITT data set cannot be easily separated from each other in LOOCV study and the classifier built from these samples cannot predict the classes of the Tumor and Patient Normal tissue samples in the Sign data set. On the other hand, the Tumor tissue samples in the PITT data set are remarkably distinct from Donor Normal tissue samples in both the PITT data set and the two independent data sets. The Non-Tumor class is a pool of both Patient Normal and Donor Normal tissue samples and therefore the performance of the classifiers based on the Tumor v Non-Tumor comparison was significantly worse than the classifier for Tumor v Donor Normal but slightly better than the classifier for Tumor v Patient Normal (regardless of classification method).

2.4.3. Patient Normal v Donor Normal classification

Based on the results above, we postulated that Patient Normal tissue samples in the PITT data set are significantly different from Donor Normal samples in the PITT data set and a classifier for Patient Normal v Donor Normal would perform successfully. Classifiers for Patient Normal v Donor Normal were built by all three methods using 63 Patient Normal and 18 Donor Normal tissue samples in the PITT data set.

The best LOOCV accuracies (Table 2.2 and Figure 2.2d) are all above 0.9 and classifiers built by BDT and WGV gave the same accuracy of 0.94. The best accuracy of 0.98 was generated by SVMs in which only two of the Patient Normal tissues were misclassified as Donor Normal and all Donor Normal tissues were perfectly identified.

Results of validating the classifiers for Patient Normal v Donor Normal on two independent data sets were also reported in Table 2.3 and Figure 2.2g, 2j. Once again, since the Singh and Welsh data sets do not have Donor Normal tissue samples, the accuracies were measured by the classification results of Patient Normal samples only. While predicting the classification of Patient Normal tissues in the Singh data set, all classifiers gave similar, very high, accuracy above 0.9. The WGV classifier achieved the best prediction accuracy with all of the 50 Patient Normal samples in the Singh data set perfectly classified. All classifiers built from the three methods gave an accuracy of 1 (100%) when validating on the 8 Patient Normal tissue samples in the Welsh data set.

These results confirm the distinction between Patient Normal and Donor Normal tissues in the PITT data set. Significantly, using the Donor Normal tissues as "baseline", the classifiers built from PITT data set predict the classification of Patient Normal tissues in the two independent test data sets very well and as such suggests the Donor Normal tissues are different from all the Patient Normal tissues we tested (see more in Section 2.4.4).

By combining these results with the results from the classifiers for Tumor v Donor Classification, we recognize that Donor Normal in the PITT data set is a unique type of Non-Tumor prostate tissues that has remarkable distinctions from Tumor prostate tissues and Patient Normal. This makes classification tasks using Donor Normal as "baseline" highly successful. It is important however to remember that these results do not imply anything about the reason that Donor Normal and Patient Normal act as distinct entities (see more in Section 2.5).

2.4.4. Classification of tissue samples from unseen classes

Classification of tissue samples from unseen classes includes: predicting the classification of tissue samples from other organ using all the classifiers we tested in this study, predicting the classification of Donor Normal tissue samples using a classifier for Tumor v Patient Normal, predicting the classification of Tumor tissue samples using a classifier for Patient Normal v Donor Normal and predicting the classification of Patient Normal tissue samples using a classifier for Tumor v Donor Normal. (The first two prediction tasks are beyond the scope of this study.) The last two prediction tasks are particularly interesting to us as they can provide insight on the specificity of the Tumor v Donor Normal and the Patient Normal v Donor Normal classifiers, when Donor Normal tissues are absent from the independent test sets.

We used the exactly same classifiers for Tumor v Donor Normal and for Patient Normal v Donor Normal in Table 2.3 to predict the class memberships of the Patient Normal tissue and Tumor tissue samples in the Singh and Welsh data sets correspondingly as those classifiers built by each method gave the best validation results. In another words, the exact parameters and reduced feature sets were used to construct the classifiers using PITT data set by each method that were then used to predict the classification of the tissue samples from an unseen class in the two independent data sets.

Table 2.4 shows the prediction results. Most of the Patient Normal tissue samples in the two independent data sets were predicted as Tumor by the classifiers for Tumor v Donor Normal built by all three methods. Whereas most of the Tumor samples were put in the Patient Normal category by the classifiers for Patient Normal and Donor Normal. This result suggest that classifiers for Tumor v Donor Normal and for Patient Normal v Donor Normal were able to distinguish tissue samples from different patient origins as both the Tumor and Patient Normal tissue samples were mostly not predicted as Donor Normal. In addition, this result provides additional support on the uniqueness of Donor Normal tissues.

Table 2.4 Results of the two types of classifiers built by PITT data set using three methods on the classification of tissue samples from unseen classes in independent data sets.

Classifiers	Tumor v Donor Normal			Patient Normal v Donor Normal		
Tissue samples for prediction in each independent data set	Patient Normal			Tumor		
Singh Data Set						
Classification methods	BDT	SVM	WGV	BDT	SVM	WGV
AC (from Table 2.3)	1	0.98	1	0.94	0.92	1
Classified as Donor Normal	1	5	5	2	0	0
Classified as Tumor	49	45	45	50	52	52
Welsh Data Set						
Classification methods	BDT	SVM	WGV	BDT	SVM	WGV
AC (from Table 2.3)	1	1	1	1	1	1
Classified as Donor Normal	0	2	1	1	1	1
Classified as Tumor	8	6	7	24	24	24

BDT: boosted decision tree; SVM: support vector machine; WGV: weighted gene voting; FP: false positive; FN: false negative; TP: true positive; TN: true negative; AC: accuracy.

2.4.5. Feature selection on classification accuracy

We have observed in Table 2.2 and 2.3 that most of the best accuracies are obtained while performing classification using a small subset of significantly expressed genes instead of the full set of genes. Comprehensive investigation of the effect of feature selection on classification accuracy results in constructing Figure 2.2 where the classification results of each classifier at different level of feature selections are plot and compared.

Performances of the classifiers built by the BDT method appear to be significantly affected by feature selection. Five of the ten results, the LOOCV accuracy of the classifier for Tumor v Non-Tumor (Figure 2.2a), the LOOCV accuracy of the classifier for Tumor v Patient Normal (Figure 2.2b), the accuracy of the classifier for Tumor v Patient Normal validating on Singh data

(Figure 2.2e) and the accuracies of the classifier for Patient Normal v Donor Normal validating on both Singh and Welsh data (Figure 2.2 g and j), show significantly increase in accuracies along with the decrease of the number of significantly expressed genes used. The accuracy of the classifier for Tumor v Patient Normal validating on Welsh data, however, earns low accuracies when decreasing the number of significant genes used. Both the LOOCV accuracy and the accuracies of the classifier for Tumor v Donor Normal validating on both Singh and Welsh data are not affected by feature selection although the best LOOCV accuracy is gained using either 20 or 50 most significant expressed genes.

The performances of the classifiers built by the SVMs, on the other hand, have not been affected much by feature selection. Only the accuracies of the classifier for Patient Normal v Donor Normal validating on Singh (Figure 2.2g) and Welsh data (Figure 2.2j) significantly increased when decreasing the number of significant genes used. All other accuracies have not shown considerable trends along with feature selection. This finding agrees with other studies[113].

Finally, the accuracies of the classifiers built by WGV showed similar trends as those of the classifiers built by support vector machines. Most of the accuracies have not been affected by feature selection. The LOOCV accuracy of the classifier for Tumor v Patient Normal (Figure 2.2b) and the accuracy of the classifier for Patient Normal v Donor Normal validating on Singh data (Figure 2.2g) decreased when reduce the number of significant genes used. The accuracy of the classifier for Tumor v Patient Normal validating on Welsh data (Figure 2.2e) increased significantly while decreasing the number of significant genes used.

In the aggregate, the BDT appears to favor small subsets of significant genes; SVMs performs well regardless of the number of genes used; whereas the WGV approach functions badly when the number of significant genes used is too small (10 genes or 20 genes). It is hard to summarize a rule to predict the perfect number of genes for a given classification method. The number of genes used to achieve the best accuracies depends on the data set and individual classification method used. Significantly, the performances of the classifiers built by the three methods are not adversely affected if the number of genes used is reduced from 12,000 to 2,000 or less. However, this feature selection dramatically decreases the complex city and cost of the classification task.

2.5.Discussion

The main goals of our work are to create highly accurate classifiers based on microarray gene expression data and make sound predictions on the classification of prostate cancer tissue samples. To this end, the classifiers for Tumor v Donor Normal and for Patient Normal v Donor Normal built from PITT data set performed very well in both leave-one-out cross validation studies and validation on independent data sets. These results were independent of the methods (BDT, SVM and WGV) used to build the classifiers. On the other hand, classifiers for Tumor v Non-Tumor (where Non-Tumor tissue samples consist of Patient Normal and Donor Normal tissue samples) and the classifiers for Tumor v Patient Normal built from PITT data set gave low accuracies in leave-one-out cross validation regardless to classification method. For comparison, the Tumor v Patient Normal classifiers were validated on independent data with the expected poor results. These results indicate that quality of data and type of baseline data employed is much more important than the specific classification method used, and is an important factor in microarray analysis of prostate cancer.

It is important to note that the Tumor v Patient Normal comparison and the Tumor v Donor Normal comparison are fundamentally different. In the Tumor and Patient Normal classification, both the Tumor and Patient Normal samples are from the same patient population (prostate cancer patients undergoing prostatectomy) and from the same type of surgical specimens (prostates with cancer taken at surgery).[38-40] Donor Normal samples are from a different population (health organ donors) and specimen types (cancer free prostates taken at the time of organ harvest)[40]. The differences picked up by the classifiers for Tumor v Donor Normal built from the PITT data set may therefore be secondary to either biology or artifact (differences in processing or population) or both. That said, the classifiers for Tumor v Donor

Normal proved capable of accurately separating Tumor from Non-Tumor samples in all three data sets containing Tumor and Patient Normal tissue samples. It is our belief that at least some of this separation may be due to biologic differences between prostate cancer and truly normal donor prostate (and, by extension, some biologic similarities between tumor and adjacent normal appearing tissue from a diseased gland). These distinctions in underlying population and tissue origins also contributed to the success of the classifiers for Patient Normal and Donor Normal built from PITT data set in predicting the classification of all the Patient Normal tissue samples in all three data sets.

We have further investigated the possible causes for the poor prediction results of the classifiers built from PITT Tumor v Patient Normal data on the Singh data set. As mentioned briefly in the results section, we built classifiers on the Singh Tumor v Patient Normal data set using all three classification methods and evaluated them through the leave-one-out cross validation and against two independent data sets (PITT and Welsh data sets). All classifiers performed very well in LOOCV (accuracy ~ 95%) and against the Welsh data set but poorly (accuracy ~ 0.72) against the PITT Tumor v Patient Normal data. The significant LOOCV differences in Tumor v Patient Normal classifiers built on the PITT (accuracy ~ 70%) and Singh (accuracy ~ 95%) indicates there may be differences in the either the population, sampling, processing or a combination between the two data sets.

These findings have implications for microarray analysis and prostate biology. Classifiers for Tumor v Patient Normal built on the Singh data using all three methods performed well in LOOCV, indicating that there were expression differences between Tumor and Patient Normal samples that can support a robust classification. Results from original Singh paper[38] also provided evidence for distinctions in gene expression between the two types of tissues in Singh data. Classifiers for Tumor v Patient Normal built on the PITT data performed poorly in LOOCV, indicating limited and fairly soft gene expression differences between sample types. Barring differences in laboratory technique or quality (which cannot be determined from the data, as no quality control data is provided) one could attempt to reconcile these differences based on potential difference in population, tumor grade or stage, or sampling. Patient Normal samples could have been taken closer or further away from the tumor. In other studies several groups have reported molecular differences in normal appearing prostate immediately adjacent to tumor[45, 170]. It is also possible that the samples were taken from different anatomic lobes (i.e. posterior versus central) and therefore might reflect topographical variation in expression patterns because of different epithelium/stromal/smooth muscle ratios.

Careful evaluation of the original papers, as well as discussions with the authors, indicates that though there were multiple, relatively small differences in samples (Table 2.1) and the way they were taken, processed, stored. It is very difficult to determine which if any of these factors are responsible for the significant difference in classifier performance. This appears to be another example for the growing consensus in the literature[93, 171] of importance of careful experimental design, tissue sampling, quality control and documentation of all aspects of microarray studies, not just statistical analysis.

Another contribution of this study came from the simultaneous application of three classification methods. This approach helped to minimize any possible impairment on the classification results because of a bad classification algorithm. The results indicate that the C4.5 based decision tree learning approach, if boosted, performed equivalently to the classification performance produced by more accepted microarray classification methods such as support vector machines and weighted gene voting.

The decision tree based classifiers are of particular interest to the clinical application of microarray analysis because, unlike many classifiers, decision tree classifiers are easy to understand and the output - the learned decision trees and the induced rules sets - are remarkably easy to interpret. Decision trees are non-parametric, can incorporate numeric and categorical attributes, and are robust in the face of missing values. Furthermore, decision trees can be used to explore and reveal correlations and interactions between genes that exist in biologic systems and provide information on the relationship between attributes and classes. The C4.5 decision tree learning method[167] also aids in the selection of a small set of relevant genes by automatically selecting genes (on the basis of information gain) that are the most informative to the classification problem at hand. This property may make decision tree learning the optimal classification method in clinical applications of gene expression data.

Despite their potential advantages, decision tree based classifiers have not been widely applied to the analysis of microarray data. Brown *et al.*(2000)[112] applied decision tree classifiers based on the C4.5 algorithm to the classification of genes and reported that results were inferior to those produced by Support Vector Machine classification. Zhang et al (2001)[115] introduced a decision tree based approach and claimed to be able to classify breast and colon cancer specimens successfully. However, the authors of that paper introduced selection bias while performing cross-validation after selection of the informative genes using the full data set. Therefore, the very high accuracies from cross validation are biased.

Finally, it is important to note that it is the underlying gene expression, a function of both biology and experimental procedure that determines the classification performance. Differences in experimental design, specimen types, patient populations between different data sets, as well as the lack of consistent quality control documentation across the entire microarray experimental process remarkably affected the microarray results and as such provided the most significant difficulties in classification tasks. .

3. CHAPTER III Integration of microarray data sets with platform dependent solutions

Presented as a Poster at Critical Assessment of Microarray Data Analysis 2003 Conference

3.1.Abstract

Comparison and integration of gene expression data generated at different institutions and across different DNA microarray platforms have become major tasks for many studies involving DNA microarray experiments. Accordingly, integration of information from several data sets of lung cancer gene expression profiling studies had become the major challenge of the Critical Assessment of Microarray Data Analysis 2003 conference (CAMDA'03). This study describes the analysis efforts attempted to integrate gene expression profiling results of lung cancer studies from two distinct research institutions using two generations of Affymetrix GeneChip® arrays. Results were presented on CAMDA'03 conference as a poster. In order to combine gene expression data, two data integration strategies were used in this study. One method integrates gene expression data using overlapped GenBank accession numbers or any other common identification from public databases across different DNA microarray platforms. The other method, in contrast, is a strategy specific for the Affymetrix GeneChip® arrays as it uses the probe-level matching information to achieve data integration. Integrated data was compared to examine whether these solutions are sufficient to achieve integrate between studies. We also discuss some of the issues relating to integrating diverse data sets.

3.2.Introduction

Gene expression profiling with DNA microarray based experiments have recently been used in molecular classification of cancers[33, 34] and prediction of clinical outcomes using clinical specimens[35, 36]. Usually, a number of studies on a particular type of cancer are conducted independently over years and/or at more than one institution because of the high cost of microarray experiments and limited availability of clinical specimens. As a result, multiple microarray data sets can be generated for the same cancer type and cover the same question in cancer biology. If integrated, these data sets could contribute significantly towards inter-study validation and the development of cancer biomarkers.

However, integration of microarray gene expression data is not straightforward. Microarray gene expression data sets may be generated from different microarray platforms or different versions of arrays within the same platform. As reviewed in Chapter I, there exist a number of microarray platforms differing in probe deposition methods, number of probes per target, probe sequences and targets identified[2, 4, 25, 28]. Even within a single platform such as Affymetrix GeneChip® arrays, various versions of arrays or array sets differ in probe and target sequences making comparison between versions difficult. Consequently, the integration of information from different microarray dataset becomes a major challenge in many analyses using data from microarray experiments.

HG_U95Av2 and HuGeneFL arrays were used to generate lung cancer data sets at two institutions (Table 3.1)[33, 35]. Both of them survey gene expression in human genomes; there are several major differences between them. First, the two arrays have different number of probe sets which represents different full-length genes. HuGeneFL is an early generation array which has 7,129 probe sets representing 5,600 full-length human genes. Most of the full-length genes on this array were selected from UniGene build 18 and the rest of them were from either GenBank or TIGR (The Institute for Genomic Research). HG_u95Av2, in contrast, has 12,600 probe sets and monitors the expression level of approximately 10,000 full-length genes, all of which were selected from UniGene build 95. There is some overlap of the full-length genes represented by the two arrays. The degree of overlapped is discussed in Section 3.4.1. Second, the number of probe pairs in a probe set on each array is different. HuGeneFL uses 20 probe pairs in a probe set to represent a transcript in the human genome while HG_U95Av2 uses 16 probe pairs instead. Furthermore, for a probe set representing the same transcript in human genome on the two arrays, the probes in this probe are not identical. These three major differences should be taken into consideration when attempting to integrate gene expression data generated from the two arrays.

In this study, we have attempted to devise two strategies to integrate two lung cancer data sets generated using either HuGeneFL or HG_U95av2[33, 35] and examine whether these solutions are sufficient to achieve integrate between studies.

3.3.Materials and methods

3.3.1. Data sets

Two data sets on lung cancer were used to plan two approaches for integrating gene expression profiling results from different generations of the Affymetrix GeneChip® arrays. One data set (Harvard)[33] was generated at Harvard using Affymetrix HG_U95Av2 arrays and the other (Michigan)[35] was generated at Michigan using Affymetrix HuGeneFL arrays. Table 3.1 summarizes and compared these two studies including: the types of microarrays, patient demographic information, clinical variables, and processions presented in the original repots. Many fields in the table have no information from the original reports and therefore filled with

"NA". The ".cel" files from both data sets were downloaded for data integration and further analysis.

Table 3.1. Comparison of the two lung cancer data sets for integration.

		Harvard	**Michigan**
DNA Microarrays	Microarray types	HG_U95Av2	HuGeneFL
	No. of Transcripts on the array	12,600	7,129
	No. of full-length Genes represented	10,000	5,600
	UniGene Build No.	95	18
	Other Sequence Databases	NA	GenBank, TIGR
	Probe pairs	16	20
Demographics	Total number of patients in study	203	96
	Primary lung adenocarcinoma	127	86
	Normal	17	10
	Age (Median)	64	63.5
	Female	71	51
	Male	53	35
	Smoking	44	48
Clinical Variables	Where is normal relative to tumor	NA	NA
	% tumor cells	variable	>70%
	Met static	12	NA
	Survival (average month)	37.5	29.5
	P53 accumulation	NA	See paper
	K ras mutation	NA	See paper
	Stage I	73	67
	Stage III	10	19
	Average Diameter	Variable	NA
	Classification	Variable	Variable
Processing Conditions	Surgery	Variable	NA
	Processing condition	NA	NA
	RNA quality Control Measure	NA	NA
	Hybridization and scanning conditions	Replicate	NA

3.3.2. Data integration strategies

Two strategies were used to integrate data from the studies since the HGU95Av2 and HuGeneFL arrays contain different probe sets. Figure 3.1 schematically depicts the two methodologies.

3.3.2.1. Integration based on GenBank accession number

In the first approach (Method I) the two datasets were analyzed using Microarray Analysis Suite version 5.0 (Affymetrix Inc. Santa Clara, CA) and then integrated by matching the GenBank accession number of each probe set in the two arrays (Fig 3.1a). Only those transcripts with GenBank accession numbers represented on both array types were used for further analysis.

3.3.2.2. Integration using the overlapped probes

The second integration approach (Method II) integrates data by using probe-level matching information provided by Affymetrix (Fig 3.1b). Many probe sets in one array type, e.g. HuGeneFL, may share/match probes with a probe set on the other array type. Only probe sets sharing probes on both array types were used for data analysis and normalization with MAS5.0. The data sets obtained afterward has been integrated as they contain only gene expression data from probe sets with matched probes. These data were then subject to further analysis.

Figure 3.2 presented details on how the integration was done by using overlapped probes. The probe matching information is provided in the Array Comparison File by Affymetrix Inc. (Santa Clara, CA). This file contains details about probe sets which shared matched probes on the two generations of Affymetrix GeneChip® arrays. For example, as depicted in Figure 3.2, the probe set A28102_at on the HuGeneFL array shares three matched probes with the probe set 31726_at on the HG_U95Av2 array. The maximum number of matched probes is 16 because typical probe sets on the HG_U95Av2 array comprise 16 probe pairs and typical probe sets on the HuGeneFL array contain 20 probe pairs. Therefore, even at the maximum level of overlap some of the probes of a probe set on the HuGeneFL array do not match any probe of the overlapped probe set on the HG_U95Av2 array. On the other hand, not every probe sets on either of the two arrays has a match on the other array.

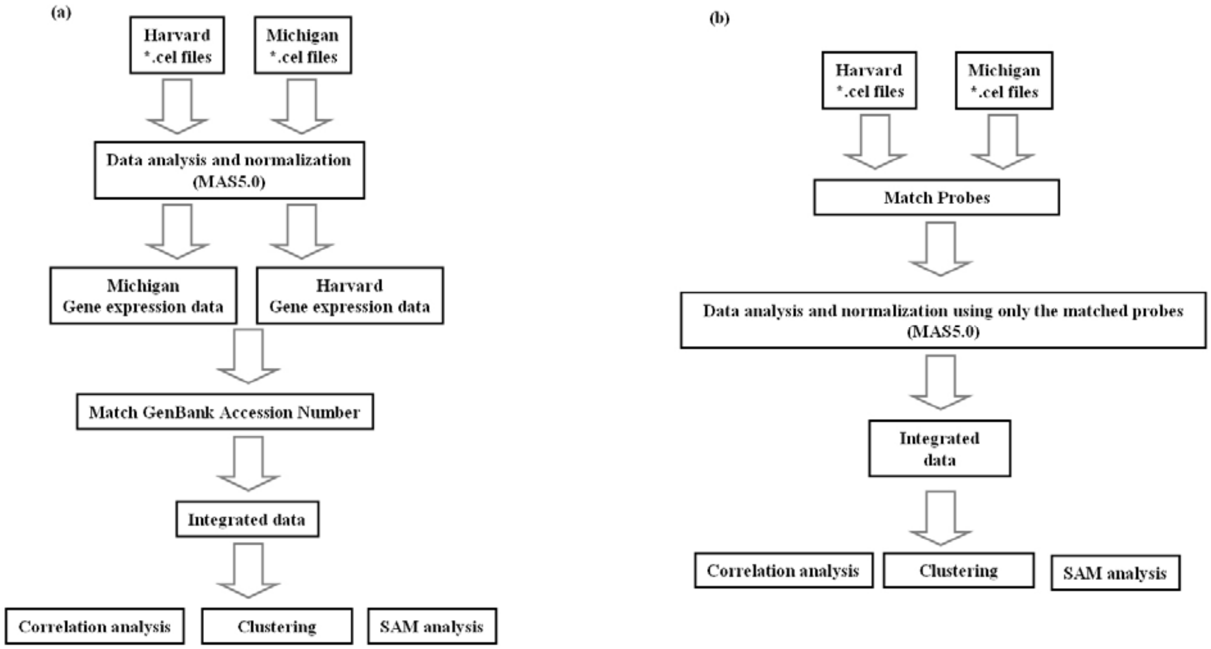
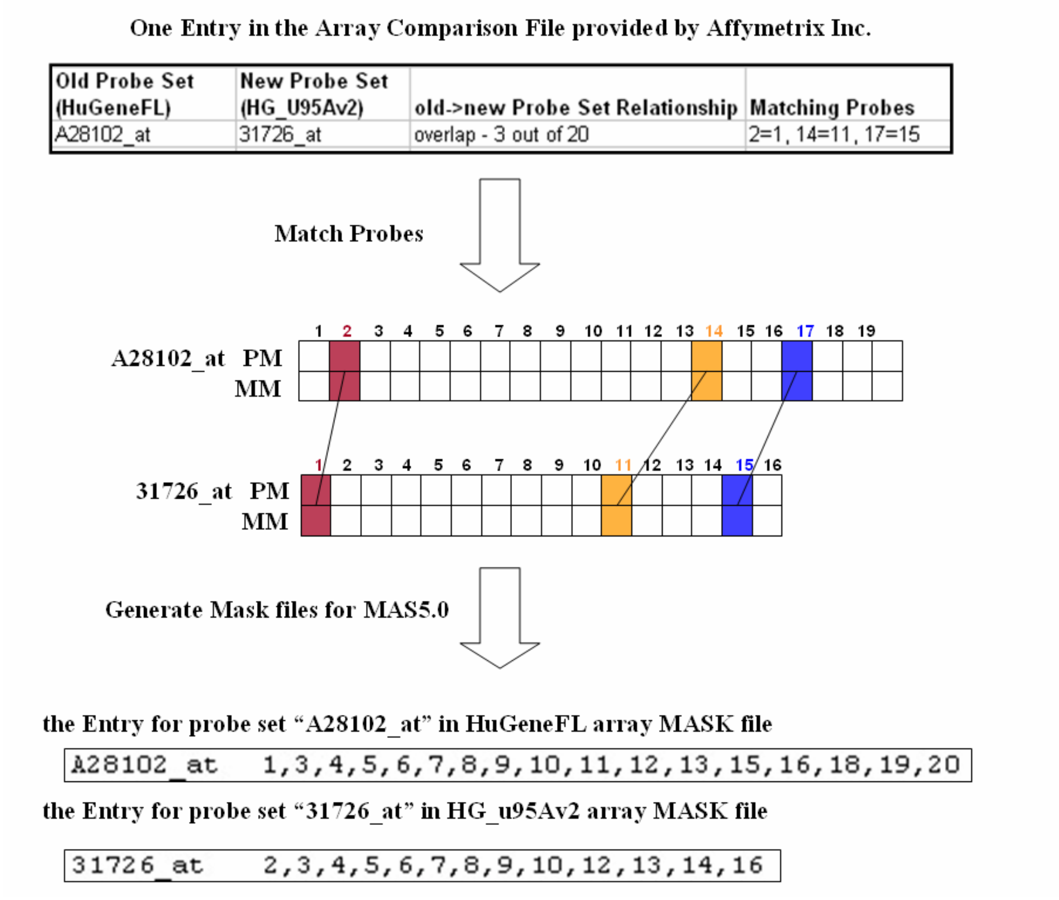


Figure 3.1 Schematic representation of the two integration strategies.

(a) Integration using GenBank accession number; (b) integration with probe-level matching information.

In order to accomplish integration, matched probes were first identified for a probe set in one array and its corresponding probe set with which it shares probes on the other array. Only the matched probes participated in the calculation of expression intensity values for the two probe sets; all other probe pairs with no matches were eliminated from data analysis by MAS5.0. If a probe set had no match at all, it was eliminated from data analysis. MAS5.0 software uses a “MASK” file to identify which probes for each probe set need to be eliminated from analysis

[172]. Therefore, after identifying matched probes, for each probe set, an entry listing the probes for elimination was inserted into the MASK file of the particular array it belongs to (Figure 3.2).


One Entry in the Array Comparison File provided by Affymetrix Inc.

Old Probe Set (HuGeneFL)	New Probe Set (HG_U95Av2)	old->new Probe Set Relationship	Matching Probes
A28102_at	31726_at	overlap - 3 out of 20	2=1, 14=11, 17=15

the Entry for probe set "A28102_at" in HuGeneFL array MASK file

```
A28102_at   1,3,4,5,6,7,8,9,10,11,12,13,15,16,18,19,20
```

the Entry for probe set "31726_at" in HG_u95Av2 array MASK file

```
31726_at    2,3,4,5,6,7,8,9,10,12,13,14,16
```

Figure 3.2 Integration using overlapped probes.

3.3.3. High-level data analysis

Using either integration approach, the two data sets were combined to create one larger data set. Hierarchical clustering[98] was applied on the integrated data. Within the integrated data set, Pearson's correlation coefficients were calculated between data from Harvard and data from Michigan. Significance Analysis of Microarray (SAM)[109] was then used to determine if after integration data from Harvard and Michigan exhibit similar expression profiles by comparing differential gene expression of tumor versus normal samples and Stage I versus Stage II tumors.

3.4. Results

3.4.1. Integrated data sets

Method I gave 5,987 overlapped transcripts across the two array types which account for 83.98% of all transcripts on the HuGeneFL array (7,129) and 47.52% of all transcripts on the HG_U95Av2 array (12,600) respectively. Alternatively, method II yielded 6,167 overlapped transcripts corresponding to 86.51% of transcripts on HuGeneFL and 48.94% of transcripts on HG_U95Av2. Of these overlapped transcripts detected by either methods I and II, 4,671 of them were revealed by both methods, corresponding to 65.52% of all transcripts on the HuGeneFL array and 37.07% on the HG_U95Av2 array. Expression data from the 5,987 transcripts from method I and the 6,167 transcripts from method II were used for further analysis. In addition, several genes that were found to be significantly expressed in the original papers were eliminated by data integration (data not shown).

3.4.2. Correlation coefficients of the integrated data

Using integrated data from method I (5,987 transcripts), Pearson's linear correlation coefficients of Harvard normal vs. Michigan normal, Harvard tumor vs. Michigan tumor, Harvard stage I tumor vs. Michigan stage I tumor, and Harvard stage III tumor vs. Michigan stage III tumor were 0.762, 0.805, 0.816, and 0.751 respectively (Table 3.2). The average of these Pearson's correlation coefficients is 0.78. Expression data of Harvard stage III tumors vs. Michigan stage III tumors had the least correlation indicating the stage III tumor samples from the two studies may be significantly different. Using integrated data from method II, these correlation coefficients were improved significantly. The average Pearson's correlation coefficient is 0.91. All four correlations reported were above 0.88 and the Harvard Stage III tumor vs. Michigan stage III tumor still gave the lowest correlation coefficient of 0.884 (Table 3.2).

Table 3.2 Pearson's correlation coefficients within integrated data.

Integration Methods	Harvard vs. Michigan				
	Normal vs. Normal	All Tumor vs. All Tumor	Stage I Tumor vs. Stage I Tumor	Stage III Tumor vs. Stage III Tumor	Average
Method I (5,897)	0.762	0.805	0.816	0.751	0.784
Method II (6,167)	0.896	0.936	0.937	0.884	0.913

3.4.3. Hierarchical clustering of the integrated data

Using the integrated data set from either method I or method II to perform clustering with all profiles from the two original studies, the Harvard and Michigan experiments separated into distinct clusters suggesting that some major underlying inter-institution variation in the two data sets cannot be eliminated by either integration methods used in this study. These variations are more significant than biological difference between tissue samples and therefore the major driving force for clustering is the inter-institution variation. On the other hand, within the clusters formed for profiles from each institute, tumor and normal samples were separated to distinctive clusters indicating that the biological difference was preserved even though different integration strategies were used (data not shown).

3.4.4. Ability to detect differentially expressed transcripts after data integration

After data integration using either method, the individual data sets, Harvard data set and Michigan data set, were used to determine if they exhibit similar expression profiles. Differential gene expression of tumor versus normal and Stage I tumor versus Stage III tumor were detected by SAM analysis. For each comparison, the lists of differentially expressed transcripts from either data set were compared and the number of overlapped transcripts for the 100, 50, 25, 10, and 5 most up- and down-regulated transcripts were listed in Table 3.3.

Using either integration methods, for the comparison of tumor vs. normal, the maximum proportions of overlapped transcripts in all transcripts in the comparison was detected when comparing the 100 most up- and down-regulated, totally 200, transcripts. As the number of differentially expressed transcripts in the list decreases, the percentages of overlapped transcripts were also decreases. In another word, when increases the stringency of SAM analysis, the number of overlapped transcripts detected from integrated data sets decreases.

Table 3.3 The number of differentially expressed genes overlapped in each comparison using integrated data from either method.

Comparisons Within the integrated data sets		Methods	# of Overlapped Differentially Expressed Genes (# of up genes / # of down genes)				
			Top 100	Top 50	Top 25	Top 10	Top 5
Harvard Tumor vs. Normal	Michigan Tumor vs. Normal	I	80/200 (40%)	36/100 (36%)	14/50 (28%)	4/20 (20%)	3/10 (30%)
		II	97 /200 (48.5%)	39/100 (39%)	13/50 (26%)	3/20 (15%)	1/10 (10%)
Harvard Tumor: Stage I vs. Stage III	Michigan Tumor: Stage I vs. Stage III	I	9/200 (4.5%)	2/100 (2%)	0	0	0
		II	5/200 (2.5%)	2/100 (2%)	1/50 (2%)	0	0

For the comparison of stage I tumor vs. stage III tumor, however, the lists of differentially expressed transcripts from Harvard data and Michigan data detected by SAM analysis were very different from each other. The gene lists did not overlap at all when the 10 or 5 most up- and down-regulated transcripts were compared regardless of integration methods. Only several genes were overlapped when the 100 most up- and down-regulated transcripts (totally 200) were compared. These results indicate that although the Harvard and Michigan data sets demonstrate some level of overlap in differential gene expression, they are not identical.

3.5.Discussion

Results from this study demonstrated that integration of information from different datasets could be accomplished to some degree. Moderate to high correlation (0.75~0.94) can be achieved between gene expression profiles from the two original studies after integrating data from two generations of the Affymetrix GeneChip® arrays with either strategy proposed in this study. However, both the hierarchical clustering results and the low levels of concordances in detecting differentially expressed transcripts after data integration indicated the existence of major inter-institution variations which cannot be eliminated or controlled by integration strategies used. The inter-institution variation may be due to differences in patient demographics, tissues, sampling methods, experimental and analysis methods (Table 3.1). It is these potential sources of variation that must be addressed in future genomics and proteomic studies to allow inter-study comparisons and to produce high quality, highly annotated data sets for biomarker validation. Therefore, the complete integration of microarray gene expression datasets cannot be accomplished until all the variations in the process of microarray gene expression analysis have been identified and well controlled.

4. CHAPTER IV *In vitro* transcription labeling methods contribute to the variability of gene expression profiling with DNA microarrays

Submitted to Journal of Molecular Diagnostics

4.1.Abstract

Considerable variation in gene expression data from different DNA microarray platforms has been demonstrated. However, no characterization of the source of variation arising from labeling protocols has been performed. To analyze the variation associated with T7-based RNA amplification/labeling methods, aliquots of the Stratagene Human Universal Reference RNA were labeled using three eukaryotic target preparation methods and hybridized to a single array type (Affymetrix U95Av2). Variability was measured in yield and size distribution of labeled products, as well as in the gene expression results. All methods showed a shift in cRNA size distribution, when compared to un-amplified mRNA, with a significant increase in short transcripts for methods with long IVT reactions. Intra-method reproducibility showed correlation coefficients >0.99, while inter-method comparisons showed coefficients ranging from 0.94 to 0.98 and a nearly two-fold increase in coefficient of variation. Fold amplification for each method was positively correlated with the number of present genes. Two factors that introduced significant bias in gene expression data were observed: a) number of labeled nucleotides that introduces sequence dependent bias, and b) the length of the IVT reaction that introduces a transcript size dependent bias. This study provides evidence of amplification method dependent biases in gene expression data.

4.2.Introduction

Analysis of gene expression with DNA microarrays has allowed reclassification of tumors based on unique molecular profiles with potentially important prognostic and therapeutic implications[94, 107]. However, there are still significant hurdles for gene expression profiling to achieve routine acceptance within the clinical laboratory. A frequent criticism for the routine clinical use of this technology is the lack of concordance among results obtained using different array platforms[173].

While it is believed that the major causes for platform-dependent differences in gene expression are due to variations in array design, probe deposition, probe sequence and gene annotation, very little attention has been paid to the bias introduced by the amplification and labeling reactions of different manufacturers. Linear, high fidelity amplification is critical as it ensures accurate replication of the size, distribution, and complexity of the initial mRNA population. Several studies have suggested that systematic biases are introduced by variations in amplification technique which could impact expression results regardless of the choice of array platform[64, 69]. These results challenge the common underlying assumption that representation of transcripts in a sample remains unchanged by the amplification and labeling protocols used prior to hybridization.

The most widely used RNA amplification and labeling technique presently in use is the T7-based method developed by Gelder and Eberwine[61]. A growing number of T7 based amplification systems are now commercially available and most incorporate modifications from the original technique. The goal of the present study was to specifically test the effect of variations in amplification and labeling protocols on gene expression results. To achieve this goal, we compared three widely used, commercially available target amplification methods[2, 73] to delineate the variation introduced by each one, and determine its potential impact on gene expression data.

4.3.Materials and methods

4.3.1. RNA sample

The Universal Human Reference (UHR) RNA (Stratagene Corp. La Jolla, CA) was used for all amplification reactions. Aliquots of the total RNA samples were prepared according to the manufacturer's protocol. Quality of the RNA was assessed by OD260/OD280 in a ND-1000 Spectrophotometer (Nanodrop Technologies, Wilmington, DE) and by capillary electrophoresis with the Agilent 2100 Bioanalyzer (Agilent Technologies, Inc. Palo Alto, CA). Purification of mRNA was performed with the Oligotex Direct mRNA Mini Kit (Qiagen Inc. Valencia, CA) as suggested by the manufacturer.

4.3.2. Target preparation methods

Methods compared in this study will be described briefly in this section. For details readers are referred to the manufacturer's manuals and selected references[2, 73, 174-176]. Table 4.1 summarizes the major differences and similarities among the three target labeling kits.

a. Affymetrix Eukaryotic Target Preparation

Two *in vitro* transcription labeling kits compared in this study are used to prepare biotin labeled cRNA targets for Affymetrix GeneChip® arrays: the Enzo BioArray High Yield^{TM} RNA Transcript Labeling Kit (Enzo) and the GeneChip® Expression 3'-amplification Reagents for IVT labeling (Affy). For first and second-strand synthesis, these two methods utilize reagents from Invitrogen Inc., and follow the same experimental steps. Hence, major distinctions between the two methods exist in the *in vitro* transcription (IVT) step. Twelve UHR RNA aliquots were labeled by each of the two methods and five were hybridized to arrays for each method. We also performed additional experiments using a modified version of the Affy method where IVT reactions incubated for only 4 hours at 37 ^{o}C (Affy4h).

Table 4.1 Comparison of the target amplification and labeling methods.

	CodeLink	Affy	Affy4h	Enzo
Starting Total RNA	1 µg	1 µg	1 µg	5 µg
Reverse Transcription Reagents	CodeLink	----------------------- Invitrogen ----------------		
RT incubation	2 hours	1 hours		
2nd-strand cDNA synthesis Reagents	CodeLink	----------------------- Invitrogen ----------------		
2nd strand incubation	2 hours	2 hours		
Biotinylated Ribonucleotides	Biotin-11-UTP	Biotin-conjugated Uridine analog		Biotin-CTP Biotin-UTP
***In vitro* Transcription Incubation**	14 hours	16 hours	4 hours	4 hours
Purification and Fragmentation Reagents	CodeLink	----------------------- Affymetrix ----------------		

a.1 First-Strand and Second-Strand cDNA Synthesis--All reagents are from Invitrogen Corp, (Carlsbad, CA) unless otherwise specified. Recommended amounts of total RNA (Table 4.1) in 8 µL Nuclease-free water were spiked with 2 µL diluted poly-A RNA control (Affymetrix, Santa Clara, CA), then incubated with 2 µL of 50 µM T7-Oligo (dT) $_{24}$ primer (Affymetrix, Santa Clara, CA) at 70^{o}C for 10 minutes and cooled on ice. Poly-A RNA controls were diluted to appropriate concentrations immediately before performing the experiment in order to maintain the same proportionate final concentration of the spike-in controls to the total RNA. First-strand cDNA was synthesized by adding 4 µL 5X 1st-strand buffer, 2 µL 0.1M DTT, 1µL 10mM dNTP, 1µL Superscript II reverse transcriptase and incubating at 42 °C for one hour. Second-strand cDNA was synthesized by adding 91µL of Nuclease-free water, 30 µL 5X 2nd-strand buffer, 3µL 10mM dNTP, 1 µL E. coli DNA ligase, 4 µL E. coli DNA polymerase I, 1 µL RNase H and incubating at 16°C for two hours. 2 µL T4 DNA polymerase was added and the reaction was incubated at 16°C for 5 minutes. Reactions were stopped by adding 10 µL 0.5 M EDTA. Double-stranded cDNA was purified using the Sample Cleanup Module (Affymetrix, Santa Clara, CA).

a.1.1 Synthesis of Biotin-labeled cRNA with the Enzo kit--Purified double-stranded cDNA was used in the *in vitro* transcription reaction using the Enzo BioArray High Yield™ RNA Transcript Labeling Kit (Affymetrix, Santa Clara, CA) at 37°C for 4 hours in a 40 µL reaction volume, containing 4 µL of 10X HY reaction buffer, 4 µL 10X biotin-labeled ribonucleotides, and 4 µL 10X DTT, 4 µL 10X RNase inhibitor mix, 2 µL 20X T7 RNA polymerase and variable amounts of RNase-free water.

a.1.2 Synthesis of Biotin-labeled cRNA with the Affy kit--Purified double-stranded cDNA was used in the *in vitro* transcription reaction using the GeneChip® Expression 3'-amplification Reagents for IVT labeling kit (Affymetrix, Santa Clara, CA) at 37 °C for 16 hours in a 40 µL reaction volume, containing purified ds-cDNA, 4 µL of 10X IVT labeling buffer, 12 µL IVT labeling NTP mix, 4 µL IVT labeling enzyme mix and variable amount of RNase-free water. Ten additional labeling reactions incubating for only 4 hours were also performed (Affy4h method).

a.2 Fragmentation and Hybridization for Enzo and Affy Protocols--One µL of purified biotin labeled cRNA was then analyzed for purity and concentration by ND-1000 Spectrophotometer and Agilent 2100 Bioanalyzer. For the cRNA prepared by Affy4h method, purified cRNA from two reactions were pooled together to achieve the required amount of cRNA for hybridization. 15 µg of purified cRNA was incubated with the adequate amount of fragmentation buffer (Affymetrix, Santa Clara, CA) at 94 °C for 35 minutes. 1 µL aliquot was used to assess complete fragmentation by capillary electrophoresis.

b. GE Healthcare CodeLink Expression System Target Preparation

Twelve biotin-cRNA samples were prepared by the CodeLink method using the CodeLink Expression Assay Reagent Kit (GE Healthcare, Piscataway, NJ). All reagents used are from this kit unless otherwise specified. 1 µg of total RNA in 8 µL Nuclease-free water were spiked with 1 µL of working solution of bacterial control mRNAs and 2 µL diluted poly-A RNA control (Affymetrix, Santa Clara, CA), then incubated with 1 µL of T7-Oligo (dT) Primer at 70°C for 10 minutes, and cooled on ice. First-strand cDNA was synthesized by adding 2 µL 10X 1st-strand buffer, 4 µL 5mM dNTP mix, 1 µL RNase inhibitor, 1µL reverse transcriptase and then incubating at 42 °C for two hours.

Second-strand cDNA was synthesized in a 100 µL reaction volume by adding 63 µL of Nuclease-free water, 10 µL 10X 2nd-strand buffer, 4 µL 5mM dNTP mix, 2 µL DNA polymerase mix,1 µL RNase H, and then incubating at 16 °C for two hours. dsDNA was purified using the QIAquik PCR purification kit (Qiagen Corp, Valencia, CA).

In vitro transcription reaction was carried out by mixing purified dsDNA with 4 µL 10X T7 reaction buffer, 4 µL T_7 ATP solution, 4 µL T_7 GTP solution, 4 µL T_7 CTP solution, 4 µL UTP solution, 7.5 µL 10mM biotin-11-UTP (PerkinElmer Corp. Wellesley, MA), and 4 µL 10X T7 enzyme mix, then incubating for 14 hours at 37 °C, final reaction volume was 40 µL. Biotin labeled cRNA products were purified with the RNeasy Mini Kit (Qiagen Corp. Valencia, CA).

15 μg of cRNA from each sample were fragmented following the recommended procedures in CodeLink target preparation manual.

4.3.3. Evaluation of amplification products

cRNA yield for all methods was assessed in a ND-1000 Spectrophotometer (Nanodrop Technologies, Wilmington, DE). Fold amplification was calculated by dividing the total cRNA yield by the estimated mRNA content (2% of total RNA) in the initial starting total RNA of each reaction. mRNA/cRNA size distribution was obtained by capillary electrophoresis with the Agilent 2100 Bioanalyzer (Agilent Technologies, Inc. Palo Alto, CA), using the "Smear Analysis" function of the 2100 Expert software (Agilent Technologies, Inc. Palo Alto, CA). Six transcript size regions: 0~0.2kb, 0.2~0.5kb, 0.5~1.0kb, 1.0~2.0kb, 2.0~4.0kb and 4.0kb~max were defined in the electropherograms and then used to determine the percentage of area under the curve for each size interval. Small amount of rRNA contaminations, both 18s rRNA and 28s rRNA, was observed on electropherograms from mRNA and cRNA from the Enzo method. rRNA proportion was subtracted from the total area under the curve and from their corresponding regions when calculating the percentage of area under the curve. However, it is important to note that size distribution in the Agilent Bioanalyzer is relative to the fluorescence intensity and does not reflect the actual number of transcripts of a given size. Four individual mRNA samples were evaluated to determine the size distribution of un-amplified transcripts.

4.3.4. Hybridization, washing, staining and data processing

Five cRNA samples from each method were hybridized to Affymetrix GeneChip HG-U95Av2 arrays which contain 12625 probe sets representing approximately 10,000 full-length genes. Briefly, 15 μg of fragmented cRNA were mixed in a hybridization cocktail with control oligonucleotide B2 (Affymetrix, Santa Clara, CA), eukaryotic hybridization controls

(Affymetrix, Santa Clara, CA), herring sperm DNA (Promega Corp. Madison, WI), Acetylated Bovine Serum Albumin(BSA) solution (Invitrogen Corp. Carlsbad, CA), 2X hybridization buffer--made from MES-free acid monohydrate(Sigma-Aldrich Corp. St. Louis, MO), MES sodium salt (Sigma-Aldrich Corp. St. Louis, MO), 5M NaCl (Ambion, Inc. Austin, TX), 0.5M EDTA (Sigma-Aldrich Corp. St. Louis, MO), molecular biology grade water, 10% Tween 20(CalBiochem, San Diego, CA), and 10% DMSO (for Affy and Affy4h methods only) and variable amounts of water to a final volume of 300 µL. 200 µL of hybridization cocktail was hybridized on each array at 37 oC for 16 hours. Each array was then washed, stained with streptavidin-phycoerythrin in a GeneChip® Fluidics Station 400(Affymetrix, Santa Clara, CA) and scanned by a GeneChip® Scanner 3000 (Affymetrix, Santa Clara, CA) as recommended by the manufacturer. Quality Control (QC) parameters were derived from the MAS 5.0 algorithm of the GCOS software (version 1.1, Affymetrix, Santa Clara, CA). Numerical gene expression data were derived from the raw intensity files using two distinct algorithms: the MAS 5.0 and the MBEI algorithm from the dChip software (http://www.dchip.org)[177]. Gene expression data will be submitted to NCBI's Gene Expression Omnibus.

4.3.5. Analysis of gene expression data

Present (P) and Absent (A) calls are based on the detection calls made by the GCOS software. For the purposes of this study, we defined that a transcript (probe set) is "truly" present in the UHR RNA if it is identified as "P" at least three times out of five replicates of any amplification labeling method.

Data from MBEI PM-only model[177] of the dChip software were used for all the transcript lists analyses. The Avadis Pride software package v3.3 (Strand Genomics, Redwood City, CA) was used for annotation, filtering, and integration of gene expression data. Michael

Eisen's Cluster and TreeView software tools (http://rana.lbl.gov/EisenSoftware.htm)[98] were used to perform hierarchical clustering and view clustering results. Coefficient of Variance (CV) for each transcript across samples was calculated by dividing the standard deviation of its intensity values over the mean and expressed as a percentage (%CV).

Two-class unpaired comparisons of gene expression data from two methods were performed with the Significance Analysis of Microarrays (SAM)[109] software tool v1.21(http://www-stat.stanford.edu/~tibs/SAM/). All gene expression profile comparisons with SAM were performed at a false discovery rate (FDR) of less than 0.03% (delta level of 3.0), except the comparison between Affy and Affy4h data, which was performed at an FDR of 0.32% (delta = 2.0). STATA software v8.01 (STATA Corp. College Station, TX) was used for all other statistical analysis including correlation studies, Mann-Whitney tests, analysis of variance and regression analysis. SigmaPlot v.8.0 (SSPS Inc. Chicago, IL) and Microsoft® Excel were used for all plots.

For each "method A to method B" comparison of intensity values with SAM, transcripts that showed significantly increased values in method A over B were labeled as "affected by A". Conversely transcripts significantly increased in method B, therefore decreased in method A, were labeled "affected by B". For the Enzo vs. Affy4h comparison, we calculated differences in Cytosine content in the target sequence of transcripts affected by these methods. The target sequence of a transcript is defined as the region interrogated by all probes in a probe set in the Affymetrix HG-U95Av2 array. Differences in cytosine content were calculated as the ratio of C to U. and expressed as G/A, thus reflecting the actual mRNA sequence. For the Affy vs. Affy4h comparison, transcript sizes reported correspond to the target mRNA sizes reported by the array manufacturer. Both transcript lengths and probe sequence information were obtained from the NetAffx website (www.affymetrix.com).

4.4.Results

4.4.1. cRNA yields

More than 30 μg of cRNA were obtained with the Affy, Enzo and CodeLink methods in almost all reactions (Table 4.2). The Affy4h method yielded approximately 10 μg on average. The CodeLink method had the highest cRNA fold amplification and showed more variability in cRNA yields, which was mostly based on lot-to-lot differences of the amplification kit (Table 4.2 Table 4.3). Lot-to-lot variability in amplification yield was not observed in the Enzo or Affy methods.

Table 4.2 cRNA yield, fold amplification, and quality control parameters from the hybridizations to HG-U95Av2 chips (Mean ±SD).

	CodeLink	**Affy**	**Affy4h**	**Enzo**	***p*-value**[§]
cRNA yield (μg)*	83.80±41.11	37.35±5.41	10.79±1.70	31.16±5.94	<0.0001
Fold Amplification*	4189.75±2055.73	1867.67±270.38	554.81±88.37	322.84±60.44	<0.0001
Median Array Intensity (raw)	128.60±46.31	109.00±10.65	150.40±11.94	207.08±47.18	0.0019
Background	44.45±8.89	50.03±2.40	66.84±4.41	77.02±18.18	0.0005
RawQ(noise)	1.57±0.28	1.76±0.13	2.18±0.09	2.53±0.49	0.0004
% of P Calls	54.50±0.70	51.20±2.80	54.60±2.30	48.40±3.30	0.0026
# of Present Genes	7476	7207	7455	6869	NA

***:** n=12. For all other rows, n=5.

[§]**:** One-way ANOVA test with Bonferroni correction.

Table 4.3 Amplification yields from each lot of the CodeLink kit.

CodeLink Kit	Lot I (n=6)	Lot II (n=6)	Lot I and II (n=12)
Yield	48.29 ± 16.59	119.30 ± 20.44	83.80 ± 41.11
Fold Amplification	2414.42 ± 829.52	5965.08 ± 1022.17	4189.75 ± 2055.73
%CV	34.35	17.13	49.06

4.4.2. Hybridization performance

All hybridizations met quality control (QC) criteria as defined by the array manufacturer; however, some significant differences were noted (Table 4.2). Compared to hybridization results from Affy and CodeLink methods, the Enzo method had statistically significant higher background (one-way ANOVA: Affy vs. Enzo *p*-value=0.005; CodeLink vs. Enzo *p*-value= 0.001), rawQ values (noise) (Affy vs. Enzo *p*-value=0.004; CodeLink vs. Enzo *p*-value= 0.001) and average median array intensities (raw) (Affy vs. Enzo *p*-value=0.002; CodeLink vs. Enzo *p*-value= 0.012).

There were no significant differences across samples in the 3'/5' ratios of GAPDH, Lys and Phe (Table 4.4). However, the 3'/5' ratios for β-Actin, *Dap* and *Thr*, were significantly higher in the samples labeled with the CodeLink method compared to Affy, Enzo and Affy4h methods (β-Actin & Thr $p<0.001$ for all methods; Dap p = 0.004, 0.006, and 0.011 for each method respectively). Interestingly, control transcripts that showed increased 3'/5' ratios are all nearly 2kb long, while the controls not affected by this bias (GAPDH, *Lys and Phe*) are all less than 1.5kb long. Additionally, rRNA sequences were detected in all but the Enzo labeling method (Figure 4.1).

Table 4.4 (3'/5') ratios (Mean ± SD) for housekeeping genes and bacterial poly-A RNA spike controls.

	CodeLink	Affy	Affy4h	Enzo	Transcript Length (kb)
Housekeeping Genes					
GAPDH	1.16 ± 0.12	1.11 ± 0.06	1.07 ± 0.10	0.91 ± 0.05	1.27
β-Actin	5.08 ± 1.73	1.44 ± 0.13	1.20 ± 0.08	1.10 ± 0.13	1.76
Poly-A RNA Spike Controls					
Lys	3.04 ± 0.24	3.16 ± 1.46	2.97 ± 1.21	2.87 ± 0.58	1.00
Phe	1.75 ± 0.22	1.65 ± 0.11	2.45 ± 0.56	2.45 ± 0.14	1.32
Dap	5.85 ± 1.12	2.34 ± 0.40	3.11 ± 0.92	2.47 ± 0.95	1.82
Thr	4.43 ± 0.36	2.02 ± 0.12	2.35 ± 0.22	2.37 ± 0.67	1.98

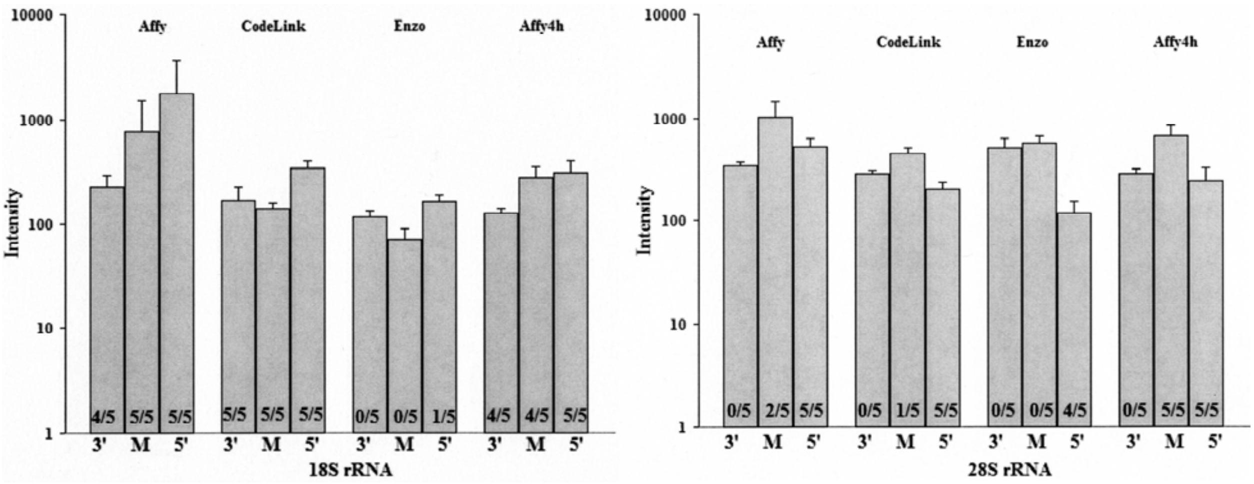


Figure 4.1. Intensity values of the 3', M, and 5' probe sets for 28S and 18S ribosomal RNAs.

The number (# / 5) on each column indicates the times this probe set is called "present" in the five hybridizations performed for each method. For example, the 4/5 of the "3'_at" probe set of18S rRNA in the Affy result (the left most bar) means that four of the five hybridizations detected this probe set as "present" in the RNA sample.

The set of present genes, as defined in the methods section, consisted of 8,281 transcripts, equivalent to 65.76% of all probe sets on a HG-U95Av2 array. The Enzo method had the lowest number of present probe sets. There was a positive correlation (R^2 =0.9553) between fold amplification and the number of present transcripts in samples from the Affy, Enzo, and CodeLink methods. Furthermore, this correlation is maintained as the stringency of the Present transcript definition goes from at least 3 of 5 replicates to 4 of 5 and 5 of 5 (data not shown). Interestingly, despite having relatively low fold amplification, the number of present probe sets in data from the Affy4h method is almost identical to the CodeLink method (Table 4.2). The four methods showed 83.3% agreement in present/absent calls for all transcripts interrogated by the HG-U95Av2 array (Figure 4.2). Of these, 6,183 (74.66%) were identified as present by all four methods. Only 2,098 were discordant between methods and, from the discordant set, less than 10% (of all transcripts on the array) were identified as present by only one method. The set of present transcripts (8,281), based on our definition in the method section, comprise transcripts both identified as present by all four methods (6,183) and discordant between methods (2,098).

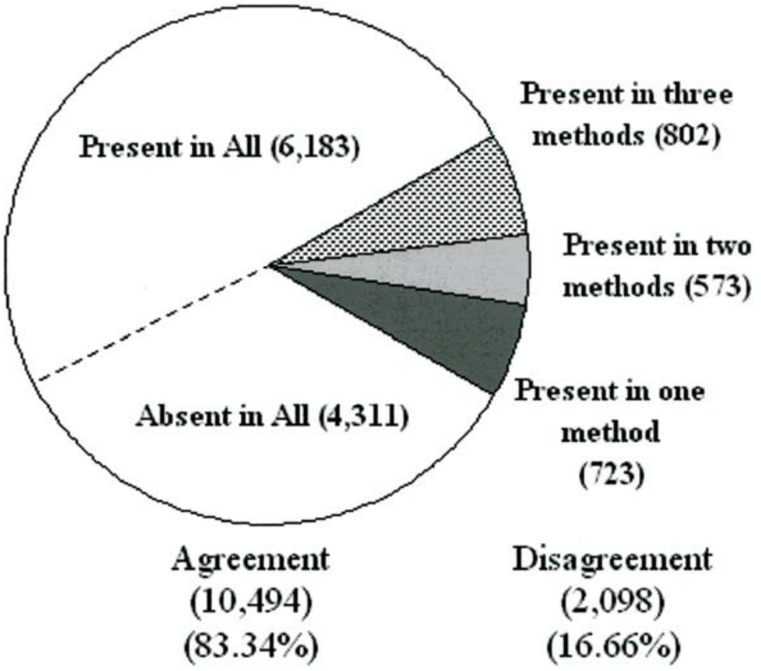


Figure 4.2 Concordance on present/absent transcript calls among the four methods studied.
Total number of transcripts in the U95v2 array is 12,592. Complete agreement among methods is represented by a white background and further divided into present and absent calls. Disagreement among methods is shaded.

4.4.3. Size distribution of cRNA products

Table 4.5 shows the distribution of cRNA products for each method. These data are derived from the electropherogram profiles of the IVT products. All methods yielded cRNA with different size distributions when compared to the non-amplified mRNA in the Universal Human Reference RNA sample with the Enzo method being most similar. The most significant difference was seen in the abundance of transcript size between 0-200bp (p <0.001) and 200 to 500 bp (p <0.001, except Enzo p =0.014). Long incubation methods (Affy-14h and CodeLink-16h) produced higher abundance of short cRNA transcripts (<1000 nucleotides), while short incubation methods (Enzo and Affy4h) produced a higher percentage of longer cRNA transcripts (>2000 nucleotides).

Table 4.5 Size distributions of mRNA in Universal Human Reference RNA and cRNA samples generated by the four labeling methods.

Base pair	UHR mRNA (n=6)	CodeLink (n=6)	Affy (n=4)	Affy4h (n=10)	Enzo (n=6)
0~200	4.2 ±1.35	11.5 ± 1.03	10.4 ± 0.52	8.2 ± 0.97	6.9 ± 0.46
200~500	4.5 ± 1.02	15.4 ± 1.30	14.0 ± 1.39	11.2 ± 1.06	7.1 ± 0.62
500~1000	17.6 ± 9.11	26.3 ± 1.92	21.8 ± 1.53	21.4 ± 0.89	16.4 ± 0.85
1000~2000	28.3 ± 10.40	25.6 ± 1.16	23.4 ± 1.22	25.1 ± 0.57	24.7 ± 0.96
2000~4000	29.2 ± 13.38	15.1 ± 1.45	19.4 ± 1.04	22.2 ± 0.96	28.4 ± 0.60
4000~max	16.0 ± 7.54	6.12 ± 1.88	11.08 ± 3.01	11.9 ± 1.77	16.5 ± 1.13

Note: The transcript abundance in each region is represented as its percentage to the total distribution.

4.4.4. Reproducibility of gene expression measurements

Pair-wise Pearson correlation coefficients of normalized gene expression measurements, within and between methods, were calculated using the set of present transcripts. Gene expression data showed excellent intra-method reproducibility and sensitivity, with correlation coefficients >0.990 for all methods (Table 4.6). The Affy and Affy4h methods had the highest inter-method correlation coefficient (*r* = 0.989), while the Enzo and CodeLink data correlate with each other the least (*r* = 0.949). With unsupervised hierarchical clustering, the arrays formed distinct clusters based on target preparation methods confirming that inter-method variability is greater than intra-method variability (data not shown).

Table 4.6 Intra- and inter-method pair-wise correlation coefficients.

	Affy	CodeLink	Enzo	Affy4h
Affy	0.9958			
CodeLink	0.9795	0.9916		
Enzo	0.9650	0.9494	0.9953	
Affy4h	0.9890	0.9703	0.9662	0.9962

4.4.5. Variability of gene expression measurements

Coefficients of variance (CV) for each present transcript were calculated across all replicates within a method (intra-assay) or across all four methods (inter-assay). As seen in Figure 4.3 a, all methods had average CVs of less than 12%, with Affy having the highest (10.45 ± 6.64%) and Affy4h the lowest (7.41 ± 4.81%). Inter-method variability was almost double of the intra-method (mean = 19.93 ± 9.87%). Figures 4.3 b-d show examples of the variability seen between methods for selected transcripts. CV plots for all transcripts in each method are presented in Figure 4.4. As has been shown in other studies, variability was higher in the low intensity region[85, 119].

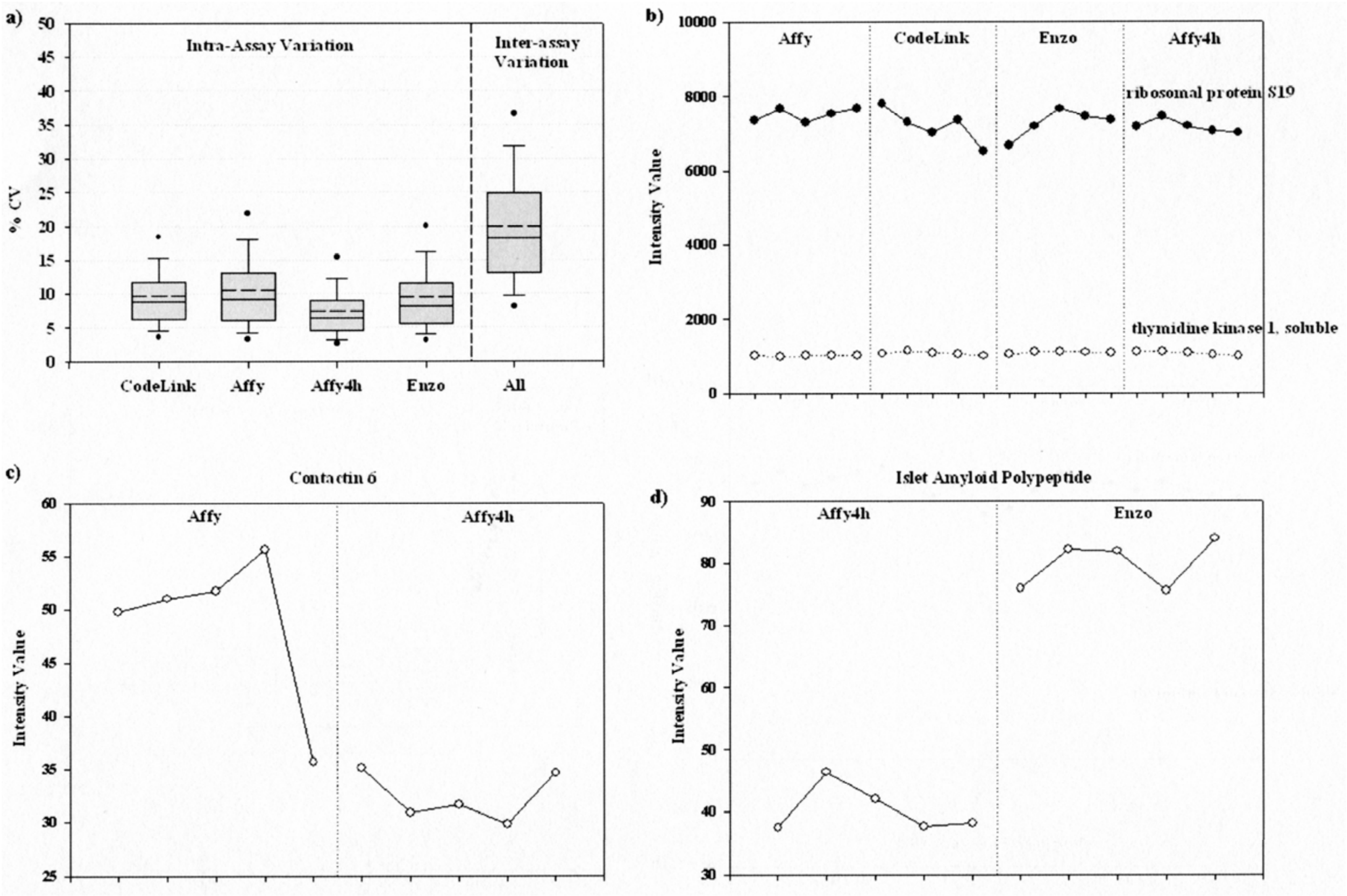


Figure 4.3 Variability in gene expression data.

Figure 4.3 legend: a) Intra- and inter-assay %CV for all present transcripts. The solid line on each box represent the median %CV while the dashed line represents the mean %CV. b) example of two transcripts with high intensity values in hybridization result showing no change across methods. c) example of one transcript with low intensity values showing difference between Affy vs. Affy4h comparison. d) example of low expressor affected by the Enzo method in the Enzo vs. Affy4h comparison.

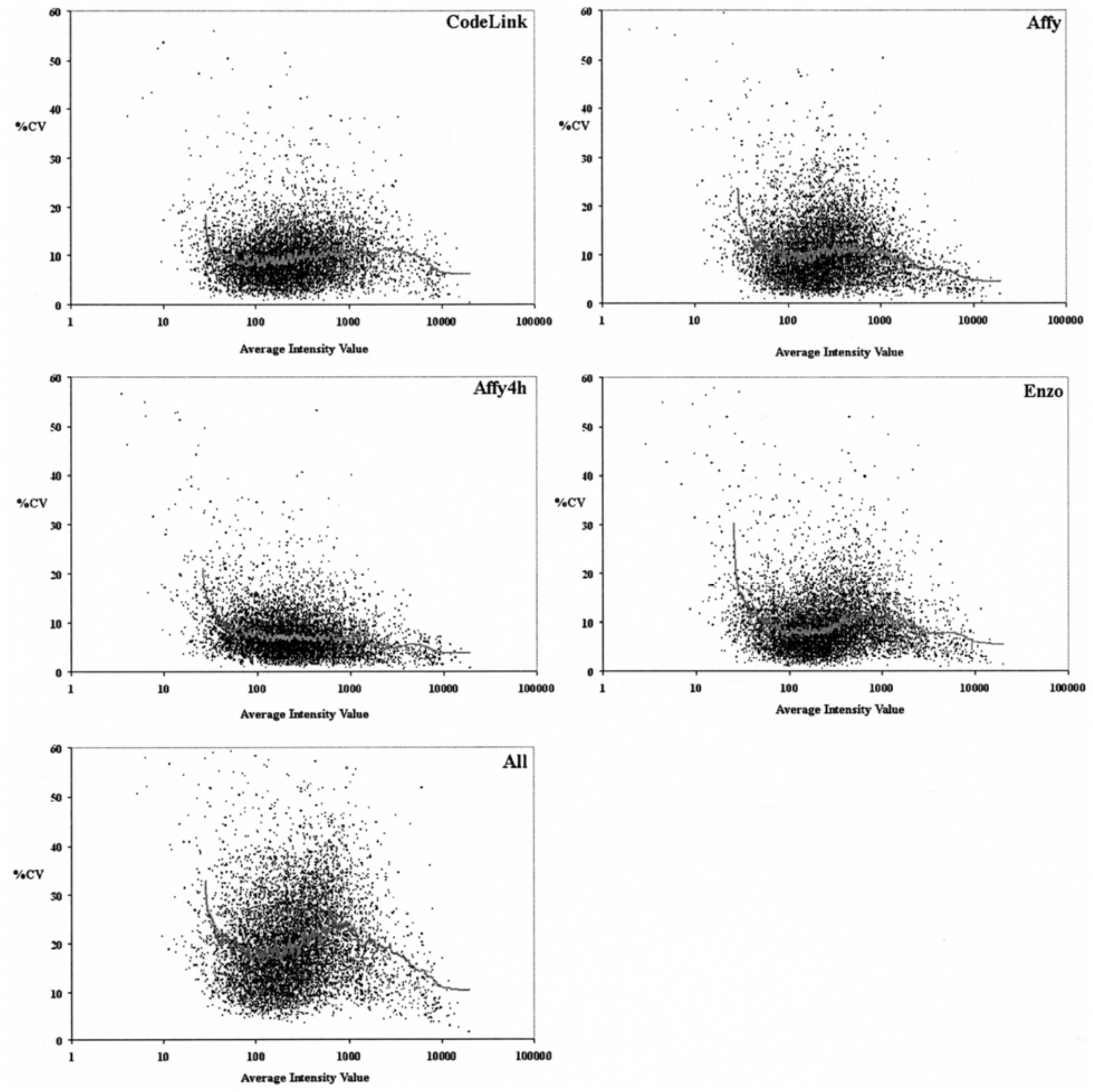


Figure 4.4 Intra- and inter-assay variations among all methods studied.

Average %CV data were plotted as a function of log average intensity value for each present transcript. In each plot, every black dot represents a transcript; a trend line (grey) depicts the moving average of %CV of every 100 transcripts. For better visualization values >60% CV are not shown.

Paired comparisons between all methods with the SAM algorithm revealed significant changes in transcript measurements, showing that cRNA targets prepared by the four studied methods have significant, reproducible, and consistent differences (Table 4.7). Since all experiments started with the same total RNA and were hybridized to the same array type, these differences are introduced by the target preparation (amplification) method. For each "method A to method B" comparison of intensity values with SAM, transcripts that showed significantly increased values in method A over B were labeled as "affected by A". Conversely, transcripts significantly increased in method B were labeled "affected by B". The comparison between Enzo and Affy4h methods had the highest number of "affected" transcripts; while the comparison between Affy and Affy4h had the lowest even at a less stringent level. For all comparisons, each method accounted for approximately half of the affected transcripts.

Table 4.7 SAM analysis results from paired comparison of all methods.

Method A	Method B	FDR (%)	Number of Affected Transcripts (%§)	Affected in A (%§)	Affected in B (%§)
Affy	**CodeLink**	0.0202	1633 (19.7)	864 (10.4)	769 (9.3)
Affy	**Affy4h**	0.3187*	2029 (24.5)	1335 (16.1)	694 (8.4)
Affy	**Enzo**	0.0109	5070 (61.2)	2577 (31.1)	2493 (30.1)
CodeLink	**Affy4h**	0.0151	3090 (37.3)	1445 (17.4)	1645 (19.9)
CodeLink	**Enzo**	0.0140	4976 (60.1)	2407 (29.1)	2569 (31.0)
Affy4h	**Enzo**	0.0082	5085 (61.4)	2585 (31.2)	2500 (30.2)

*: Delta for this comparison was set at 3.0, all others at 2.0.

Since there are multiple factors that could contribute to the observed inter-method differences, we performed two focused comparisons that allowed us to isolate the sources of variation: a) the Enzo vs. Affy4h comparison was used to analyze the effect of double nucleotide labeling, and b) the Affy vs. Affy 4h comparison was used to analyze the effect of long *in vitro* transcription reaction time. From all the methods studied, Affy and CodeLink are the most similar in terms of workflow; however, comparison between these two methods still showed affected transcripts that could not be explained by the variation sources discussed above.

4.4.6. Sources of variation

4.4.6.1. Dual labeling

The Enzo method uses double nucleotide labeling (biotin-CTP and biotin-UTP) while others use one (Table 4.1). Samples labeled with this method had higher average un-normalized fluorescence intensity values than all other methods (Table 4.2). As seen in Table 4.6 for the Enzo/Affy4h comparison, 61.4% of all transcripts have significantly different gene expression values, and are therefore affected by the method-dependent variation.

We hypothesized that if this method-dependent variation is a direct result of the double nucleotide labeling, then the transcripts that show higher gene expression values with the Enzo method will have a higher Cytosine content in the transcript sequence interrogated by the probe set, since this nucleotide is only labeled by this method. This was expressed as the G/A ratio of the target transcript sequence as defined in the Methods section. The average G/A ratio of transcripts showing elevated expression in Enzo data was 1.166 ± 0.485, which is significantly higher than those of transcripts increased by the Affy4h method (0.773 ± 0.305; Mann-Whitney test: $z = -32.477$ $p < 0.00001$). When transcripts that are affected significantly by the two methods are categorized according to their G/A ratio, we found that 93.7% of transcripts with ratios >2.0 show significantly higher values with the Enzo method and 84.70% of genes with ratios <0.5 show higher values with the Affy 4h method (Figure 4.5).

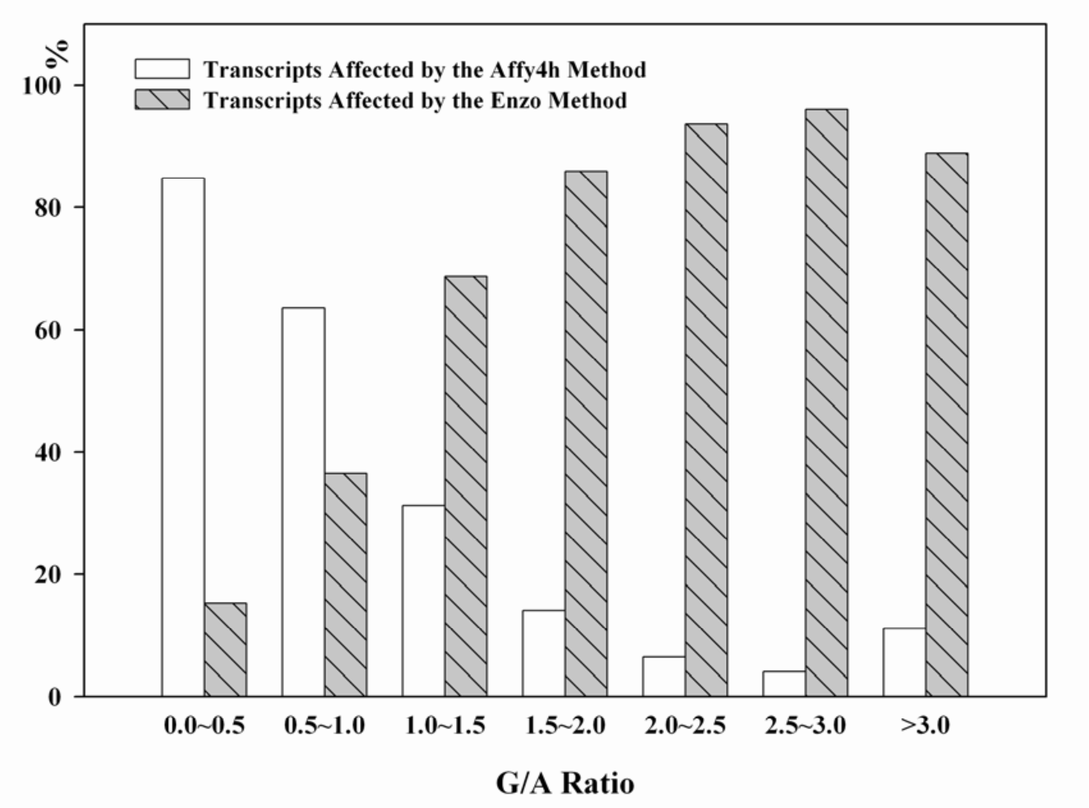


Figure 4.5 The G/A ratio of the affected transcripts from SAM analysis of the Enzo vs. Affy4h comparison.

Enzo and Affy4h affected transcripts (totally 5,085) were divided into groups based on their G/A ratios which were obtained from the target sequence information provided by array manufacture. In each group, the number of affected transcripts by each method and the corresponding percentages of total number of affected transcripts in this group were obtained. The percentage of affected transcripts of each method in each group was plotted as a function of the G/A ratio.

4.4.6.2. Incubation time

Given that the Affy and Affy4h methods only differ in the length of IVT incubation time (Table 4.1), comparison of these two methods provides an insight on how this factor affects gene expression data. In this comparison, 24.5% of all present transcripts are significantly different between Affy and Affy4h methods with a delta of 3.0 (FDR= 0.3187%).

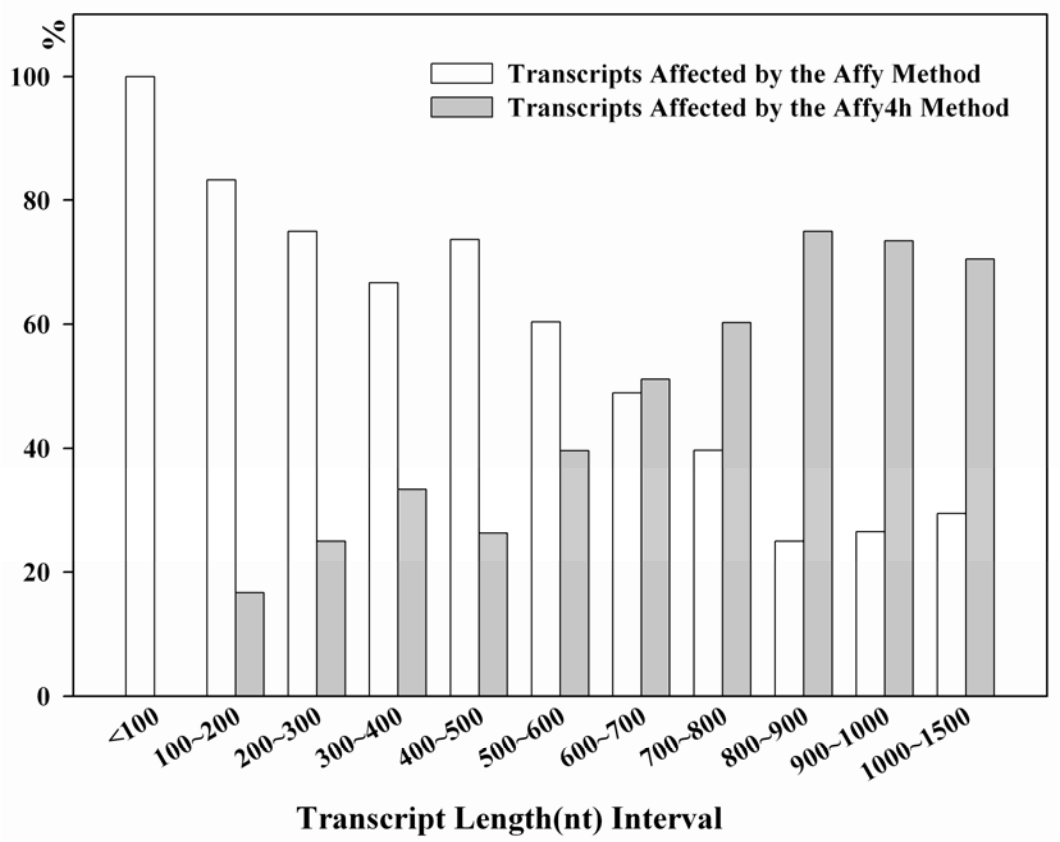


Figure 4.6 Transcript lengths of the affected transcripts from SAM analysis of Affy vs. Affy4h comparison.

Affected transcripts from SAM results were grouped based on their transcript lengths. In each interval, the number of affected transcripts and their percentage in the total number of affected transcripts in this interval were gathered. The proportion of affected transcripts of each method in each interval was plotted as a function of transcript length. Please note only the transcripts shorter than 1.5kb were of interest to this analysis.

Based on the transcript size shift observed with long IVT reactions, we hypothesized that transcripts with significant higher expression values in samples labeled with a long (overnight) IVT are more likely to be short transcripts. The analysis of 3’/5’ ratios of control genes shown above revealed that the 3’ end of transcripts > 1.5 kb was preferentially amplified by at least one of the long incubation methods (Table 4.4). Therefore, we investigated if genes < 1.5 kb would be preferentially amplified by a long-IVT labeling method. Figure 4.6 shows the percent of transcripts <1.0 kb that are selectively increased in the Affy method in comparison to the Affy4h. These data show an inverse relationship between transcript length and the percentage of transcripts whose expression values are increased by the long IVT. Linear regression analysis shows an R^2 of 0.9291, indicating a strong association between the increase of transcript length and the decrease of the proportion of long-IVT affected transcripts. This association could not be found when a comparison of both long IVT methods (Affy/Codelink) was done (Figure 4.7).

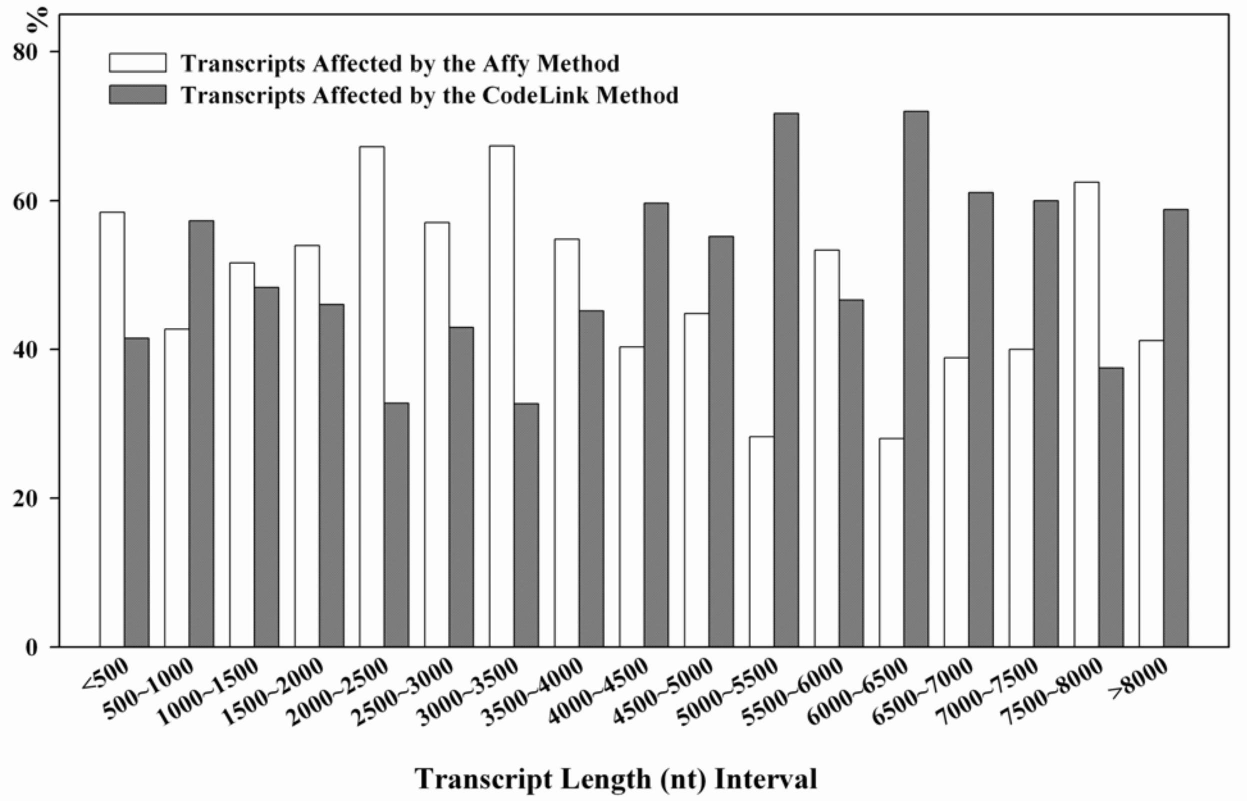


Figure 4.7 Percentage of affected transcripts by each method in the Affy vs. CodeLink comparison (long IVT), grouped by transcript length.

Data were plotted using the same strategy described in the legend for Figure 4.6 No correlation with transcript length was found in this comparison.

4.5.Discussion

This study demonstrates specific biases in gene expression data introduced by commercially available T7 RNA polymerase based amplification reagent kits and protocols. Although T7 amplification is generally regarded as linear, several studies have shown differences in gene expression between amplified cRNA (single or double round) and non-amplified mRNA[64, 69, 70, 133, 151]. Our results corroborate and extend those obtained in other studies, and show that gene expression results can show biases that are dependent on the number of labeled nucleotides in the amplification kit or in the length of IVT reaction, which translates to a transcript size-dependent bias.

Most researchers judge labeling kit for DNA microarrays based on their performance in yielding sufficient labeled cRNA for hybridization. However, our results suggest that attention should be paid to the number of biotinylated ribonucleotides used for labeling at the in vitro transcription step. When comparing single vs. double nucleotide labeling with normalized data, we found that approximately 30% of the present genes had substantially higher gene expression values in Enzo (double nucleotide) compared to Affy4h (single nucleotide), suggesting the data sets generated from methods using two labeling nucleotides are not directly comparable to data sets derived by using a single labeling nucleotide. It has been shown previously that incorporation of biotin-CTP is not as efficient as biotin-UTP.[73, 178] Our results are in agreement with these findings, since we found differences when the G/A ratio was higher than 2, indicating that at least 2 incorporated biotin-CTPs per biotin-UTP are necessary to significantly increase the amount of fluorescent signal per transcript. However, given the complexity of the labeling process, and the hybridization reaction, it is unclear if the biases introduced by the number of labeling molecules can be corrected by a normalization method.

We also demonstrate that the distribution of transcripts shifts towards shorter cRNA products in protocols with long IVT incubations, suggesting enhanced amplification of short transcripts. This is further corroborated by the fact that short transcripts were more likely to be increased in cRNA samples from long IVT labeling methods. Interestingly, Spiess and collaborators reported a similar cRNA size shift with long IVT incubation, but suggested that degradation of cRNA molecules by T7 RNA Polymerase accounted for this observation.[156] However, in our results long incubations consistently gave higher yields, which contrasts with their decrease in cRNA yield after 5h. Furthermore, in their description of exonuclease activity of T7 RNA polymerases, Sastry and Ross indicated that this activity is only unmasked in paused/arrested transcription complexes and that the kinetic balance during normal transcription was balanced towards polymerization[179]. We speculate that the degradation and/or decrease in IVT yields seen by Spiess and others [151, 156] with IVT reactions exceeding 4h, could be a result of paused transcription complexes due to depletion of reaction components. New IVT kits that are designed for longer incubation times seem to overcome this problem. Although the degree of amplification correlated with the increase in short cRNA transcripts, we were unable to assess the role of enzyme concentration between protocols with identical incubation times because the kit manufacturers would not provide this proprietary information.

In this study, the number of transcripts identified as P in a sample, was directly related to the degree of amplification achieved in all methods but one (Affy 4h). This suggests that transcripts actually present in a sample are not always amplified successfully, which contributes to the variability within and between assays. In fact, as seen in other studies[119], variability in gene expression measurements was most pronounced in the low fluorescence intensity range, i.e. in the low expressor transcript range, as would be expected if low abundance transcripts are not efficiently amplified each time. It is interesting to note that the Affy 4h method, which used pooled reactions due to low fold amplification, yielded similar P calls as the CodeLink platform, which showed the highest fold amplification. These results suggest that multiple labeling reactions may be more effective at amplifying low-expressor transcripts, because more transcription initiation events may occur with multiple short-term incubations. Further testing of this hypothesis is currently underway in our laboratory.

Intra-method variability reflects random errors created during the performance of a specific method, while inter-method variability comprises both random experimental errors and systematic biases. In the present study, all methods provided low intra-method CVs, but inter-method variability was considerably higher. Average CV across any two methods ranged from 15.65% to 20.44% approximating the average %CV across all methods of 19.93%. Other studies have reported correlation coefficients for the CodeLink and Affymetrix platforms between 0.59 to 0.79[119, 128, 180]. In our study, we obtained higher correlation coefficients between these two platforms, which could reflect the fact that all samples were hybridized to the same array type, therefore isolating only the variability contributed by the labeling method.

Another significant difference observed between labeling methods was under representation of 5’ probes from genes larger than 1.5 kb with the CodeLink method. This phenomenon was observed by Baugh et al[69], and was demonstrated to be related to inefficient reverse transcription. Indeed, when comparing the CodeLink method against all others, which share a common RT step, the former requires a longer incubation period (2h vs. 1h) that may lead to depletion of dNTPs and early termination of reverse transcription reactions yielding 5’ truncated cDNA products. It is also possible that IVT further contributes to 5’ under-representation when the T7 RNA polymerase fails to transcribe full-length transcripts. It is likely that the majority of gene expression results are not affected by this phenomenon, since most probes in current array designs are 3’ biased, but this factor should be taken into account for probes that interrogate the 5’ region of selected transcripts.

In summary, our results indicate that individual amplification methods significantly bias gene expression data, despite the fact that they are all derivatives of the T7 RNA polymerase based linear amplification. We have shown that part of this variability can be explained by: the number of biotinylated nucleotides used in the labeling reaction and the length of the in vitro transcription reaction. These biases are not corrected by intensity based normalization techniques such as the invariant set normalization method[177], and therefore can generate discordant results even with the same sample. As shown recently, concordance between different platforms has improved substantially thanks to advances in gene annotation and array design[130] and high reproducibility among laboratories can be achieved when the same protocols and array platforms are employed[121, 181]. Our results emphasize the importance of standardized target preparation methods in order to optimize gene expression analysis and achieve a consistency compatible with clinical application of this technology. These findings should be taken into account when comparing data from different platforms, and in standardizing protocols for clinical applications.

5. CHAPTER V Conclusions and future prospects

5.1.Conclusions

The goal of this thesis work is to help improve bioinformatics support and quality assurance (QA) for DNA microarray based gene expression profiling, with the goal of molecular diagnosis implementation of DNA microarrays as diagnostic and prognostic tools in clinical pathology laboratories. Rigorous quality assurance and quality control (QA/QC) have been applied in the clinical laboratories as a critical component to guarantee the delivery of high-quality test results by controlling the variance and detecting measurement errors. For DNA microarray based gene expression profiling, QA and QC can be achieved through standardizing both the technology and experimental procedures. Work in this thesis provides insight into the problems existing in the experimental procedure associate with gene expression profiling.

The most important conclusion that can be drawn from observations in this dissertation is that significant levels of variations can be introduced into microarray gene expression data either by tissue sampling or by the target preparation method and that these biases often overwhelms the most powerful statistical analysis. Variations introduced by tissue sampling have been shown to interfere significantly the accurate classification of tissue specimens from cancerous and disease free donor prostate and this principle almost certainly extends to all organ systems. For the classification of prostate tissue specimens using classifiers built on microarray data, results show the selection of the tissue baseline; normal prostate tissue specimens from prostate cancer free donors versus normal appearing prostate tissue from prostate cancer patients. In addition, results from this dissertation showed that the decision-tree learning algorithm can be successfully applied to the classification of cancer using microarray data although even the best analysis could be significantly undermined by experimental biases such as these discussed above.

DNA microarray allows the high-throughput gene expression profiling of any biological system by simultaneously surveying the expression level of tens of thousands of transcripts in massively parallel fashion and across many cellular conditions. Gene expression data from DNA microarray based experiments have both a massive data volume and exceptionally high dimensionality and, as a result, it becomes a major challenge to discovery biologically meaningful gene expression patterns from such data sets. It is not hard to understand why algorithm design and application became a major task in the application of DNA microarray technology for gene expression profiling. Few studies had been done on classification using decision-tree learning at the time the study in this dissertation was conducted. Results in this thesis show that the decision-tree learning algorithm performed as well as, if no better than, several popularly used classification algorithms on partitioning prostate tissue specimens using solely microarray gene expression profiling data. However, unlike the popular analysis methods of the time, decision-tree learning algorithm created a classifier in the form of a tree structure, which could be used to suggest potential underlying relationships between genes or potential linkage within pathways; these features have made the decision-tree learning algorithms attractive for classification tasks using microarray data. Despite the success of the decision-tree learning algorithm, however, the performance of all three classification methods was clearly impaired by the limited quality of the microarray gene expression data sets themselves.

Gene expression data from DNA microarray based experiments should delineate the composition and the relative abundance of each transcript in a transcriptome and this should be a function of the biological events happening at the time cells or tissues were harvested. However, noise and biases at various levels have been observed in microarray gene expression data and have significantly interfered with the accurate discovery of unique patterns of gene expression in a cell or tissue specimen. In many cases, the noise and bias introduced have been much larger than the biological signals themselves. This has been a major obstacle for the clinical application of DNA microarray technology as a diagnostic and prognostic prediction tool for clinical application.

Recently efforts have been focused on identifying possible sources of variations by comparing microarray gene expression profiling results generated from different microarray platforms, various institutions, and multiple laboratories. DNA microarray platform, tissue sample, laboratory and array replication are the major sources recently identified which can introduce significant levels of variations in microarray gene expression data[120, 121]. RNA labeling, hybridization, data acquisition, and data analysis methods, if standardized, have also been proven to significantly improve the reproducibility of gene expression profiling between datasets produced on different array platforms and across different laboratories[120, 121]. Significantly, however, no study has been performed to formally investigate the level and source of these variations.

T7 RNA polymerase based in vitro transcription labeling method is the most popularly used RNA labeling method. Past works have proven the linearity of this methodology and pointed out the potential for biases introduced by RNA amplification. However, no studies had systematic evaluated the significance of those biases or investigated their source. Results from the second aim of this dissertation demonstrated, for the first time, that significant levels of variations can be introduced into microarray gene expression data by several RNA labeling methods even though they are all derivatives of the T7 RNA polymerase based method. More importantly, statistically significant associations have also been established between the variations observed and their possible sources. Specifically, variations in the number of biotinylated nucleotides used in labeling have been shown to be responsible for the alteration of gene expression pattern of approximately 30% transcripts presented on a microarray. Furthermore, the incubation time of the in vitro transcription has been shown to significantly bias the gene expression pattern of short/small transcripts significantly. The observed variations/biases introduced into the experimental data set cannot be eliminated or controlled by applying advanced normalization algorithms such as the invariant set normalization, indicating data from experiments using target preparation methods with different number of labeling nucleotides or IVT reaction time may not be directly comparable.

Furthermore, although results reported were from the comparison of three particular methods/kits these observations can be generalized to other RNA labeling methods. First, these results show that the variations introduced by RNA amplification and labeling methods are significant. The average coefficient of variance of all transcripts on the array across all three methods is approximately 0.2 (Figure 4.3 and 4.4). For each paired SAM analysis, a large number of present transcripts were significantly altered/biased (20%~60% of all present transcripts on the array) (Table 4.7). Since the three methods are all derivatives from the T7 RNA polymerase based in vitro amplification approach, greater levels of variations may be expected in microarray gene expression data if RNA labeling methods used are fundamentally different from each other.

Secondly, the association between the number of biotinylated nucleotides and biased expression pattern of transcripts is not unique to the labeling methods used in this study. When comparing microarray data from RNA labeling methods that differ with the number of labeling nucleotides, results may not be directly comparable. This is because the number of labels on each target molecule (cRNA or cDNA) varies when the number of labeling nucleotides changes. Depending on DNA microarray platform used, this variation may be or may not be controllable by normalization or other algorithms. For Affymetrix GeneChip® arrays, no easy solution is expected as the probe sets are designed with high redundancy and hybridization of targets to each probe is not yet a fully understand procedure, but for arrays incorporating one probe per transcript, solutions or approximations may be made. The bottom line is that caution should be taken when attempting to compare gene expression data if the labeling methods use different number of labeling nucleotides.

Third, observations from this dissertation also suggest results from RNA labeling methods using T7 RNA polymerase based *in vitro* transcription is different and may not be comparable directly if the length of IVT reaction is not appropriately controlled. Optimum IVT reaction times are 4~5hours based on the studies on hand in this dissertation in Chapter IV.

Past work has shown that tissue sampling is an important source of variations in DNA microarray gene expression profiling. In this dissertation, results show that tissue sampling affected the performance of classifiers built using microarray gene expression data. When classifiers built on gene expression profiles of normal appearing tissues adjacent to prostate tumor and profiles of prostate tumors were not able to classify correctly prostate tumors from other institutions. On the other hand, if using the profiles of prostate specimens from prostate disease free organ donors to build the classifier instead, classifiers performed well on distinguishing prostate tumor specimens, indicating that biases induced were as great as even the most profound biologic signals. In addition, results from the attempt of integrating lung cancer data from different generations of the Affymetrix GeneChip® arrays demonstrated how variations from all sources (tissue sampling and handling, patient demographic information, experiment methods, analysis methods, etc. Table 4.1) dwarfs the “cancer” relevant biological signal, causes specimens to cluster by institution instead of biology and makes difficult, if not impossible, the integration of microarray data sets from different institutions if appropriate QA and QC are not utilized. Most importantly, these findings highlight the importance of correct/appropriate tissue sampling in applying DNA microarray gene expression profiling in cancer research and possible clinical application.

Lastly, observations and conclusions from this dissertation emphasize the importance of standardized target preparation methods and tissue sampling in order to optimize gene expression analysis and achieve a consistency compatible with clinical application of this technology. These findings should be taken into account when comparing data from different platforms, and in standardizing protocols for research and clinical applications.

5.2.Future prospects

The ultimate goal is to achieve rigorous quality control and quality assurance by standardization of both the DNA microarray technology and the experimental procedure so that good quality and comparable gene expression profiling results can be created at each individual laboratory. Moreover, these high quality data sets can eventually be shared and made available for meta-analysis. Once standardized, DNA microarray technology will become a powerful tool for research use and for diagnosis and prognosis in clinical laboratories. However, because biases induced by experimental procedure often cannot be “controlled” by later analysis, if data is to be shared or compared, standardization cannot be done in one laboratory or a single institution. Both the research community and the DNA microarray related industry should work together toward accomplishing this goal. Every result and each raw data set will contribute to the overall efforts and bring us closer to the ultimate goal of transparent sharing of gene expression data sets. Therefore, at the end of this dissertation, two types of future prospective work are proposed: (1) work towards to improving comparability of existing microarray gene expression profiling data sets, (2) work towards furthering experimental standardization of gene expression studies.

5.2.1. Developing algorithms to approximate, control or eliminate variations introduced by the number of labeling nucleotides

Many previously published studies used RNA amplification and labeling methods with two labeling nucleotides while the majority of current studies use methods with one single labeling nucleotide. The significant levels of variation due to the number of labeling nucleotides (as demonstrated in this thesis), therefore, becomes a hurdle for the comparison and integration of microarray gene expression results generated recently with data from the past two years.

A possible solution to this problem is to develop algorithms which can simulate and control for the events occurring at hybridization and, at the same time, take into account of the number of labeling molecules on a target molecule. Before such an algorithm can be developed, more analysis need to be done on the effect of the number of labeling nucleotides on other types of DNA microarray which may have different probe length and use different labeling dyes. These may all contribute to the overall variations observed across different data sets and will be used to develop important parameters in the approximation/correction algorithm.

5.2.2. Cross-platform comparison and integration of prostate gene expression data generated with standardized target preparation method

As target amplification and labeling methods have been demonstrated to introduce significant level of variations into microarray gene expression data (Chapter IV), cross-platform comparison and integration studies should be cautious in the use of multiple target preparation methods. Therefore, the data set that was generated in our laboratory using prostate tissue specimens from both tumor patients and disease-free organ donors is very useful in this work. The uniqueness of this data set is that targets for hybridization were all prepared with one RNA labeling method. Three types of arrays were used: the CodeLink oligonucleotide arrays from GE HealthCare; the HG_U95Av2 arrays and the HG_U133A arrays from Affymetrix.

The objectives of this proposed study are as follows. First, variations due to the differences between DNA microarray platforms will be measured and characterized. Three paired comparison can be made: CodeLink array vs. HG_u133A array, CodeLink array vs. HG_U95Av2, and HG_U133A array vs. HG_U95Av2 array. Previous studies reported the correlation of gene expression data from CodeLink arrays and Affymetrix GeneChip® arrays is from 0.5 to 0.79[118, 119, 128]. Most of these studies did not control for the variation introduced by target preparation methods. Results from our data sets are expected to show better concordance of gene expression data from the two platforms. These comparisons will help to investigate further the possible sources of the observed variations.

We expect to use this data set to study a data integration strategy with matched sequences and matched probes across different platforms. Previous studies from other groups have shown that cross-platform concordance of microarray gene expression data can be improved by using probes with matched sequences[182]. Variations from target preparation methods and other sources were not controlled and therefore the levels of improvement may be less than optimum. The data set we have generated is unique and valuable because it was truly with very well controlled experiments to minimize possible variations and used a single target preparation method. Comparison and integration results from this data set should reflect more closely the real level of improvement possible by using sequence matched probes.

5.2.3. Developing an optimal method for target amplification and labeling

An optimal target preparation method would preserve 100% the integrity, composition, and relative abundance of each transcript in a transcriptome. Results from this dissertation show that that the Affy 4h method, which used pooled reactions due to low fold amplification, yielded a similar number of "present" transcripts as the CodeLink platform which showed the highest fold amplification. These results suggest that multiple labeling reactions may be more effective at amplifying low-expressor transcripts, because more transcription initiation events may occur with multiple short-term incubations. Further testing of this hypothesis is currently underway in our laboratory.

5.2.4. Future works related to other sources of variation in DNA microarray based gene expression profiling

In this thesis, tissue sampling and target preparation methods have been investigated as sources for the variations in microarray gene expression results. There are many other ones left unstudied.

For example, the probe design for different DNA microarray platform may affect the microarray data if probes are not optimized to hybridize with their intended target molecules. The issues with probe design are whether redundancy should be applied and the optimal probe length with or without redundancy. Controversial results regarding these two issues have been reported recently[28, 183, 184]. Large scale, systematic studies need to be carried out to investigate this problem.

Hybridization is a critical step in microarray experiments (Figure 1.2a). Current protocols typically use 18 to 24 hours at hybridization and assume that this length of incubation is long enough for hybridization reactions to reach equilibrium. However, a study by Sartor *et al.*[185] recently carried out a study to investigate the effect of increasing hybridization time (from 18 hours to 42~66 hours) on gene expression data using two-channel long oligonucleotide microarrays. Their results show that hybridization results from prolonged hybridization yielded more genes detected as present on the array, higher signal-to-noise ratio, and better reproducibility compared to results from 18-hour hybridization. These results suggested that hybridization reaction does not reach to equilibrium, as assumed, at 18 hours and, consequently, gene expression data from such hybridization will not reflect faithfully the level of expression of transcripts in a transcriptome. Specifically, as there will be high proportion of nonspecific hybridization before reaching equilibrium, some genes presented in the transcriptome may not be detected as present in microarray data sets and the fold change of the differentially expressed genes may be underestimated. The study summarized here demonstrates the variations introduced by hybridization on a specific type of DNA microarrays. It is likely that variation from the length of hybridization is not unique to the two-channel long oligonucleotide array used in that study. Other studies had also presented some preliminary results support this observation[23]. However, few studies thoroughly investigate the extent and source of variations introduced by hybridization. Furthermore, variations associated with hybridization time may vary with target concentration. Therefore, target concentration should also be taken into account when such studies are designed.

In summary, we have proposed several possible prospective studies to extend the observations and results from this thesis. These future studies will help to improve standardization of experimental procedures, provide better integration of microarray data sets which have been generated, and improve our understanding of the sources of variation in DNA microarray data.

APPENDIX A

Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors

BMC Cancer

BioMed Central

Research article

Open Access

Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors

Uma R Chandran, Rajiv Dhir, Changqing Ma, George Michalopoulos, Michael Becich and John Gilbertson*

Address: From the Department of Pathology, University of Pittsburgh, Pittsburgh, PA 15232, USA

Email: Uma R Chandran - chandran@pitt.edu; Rajiv Dhir - dhirr@msx.upmc.edu; Changqing Ma - chmst40@pitt.edu;
George Michalopoulos - michalopoulosgk@msx.upmc.edu; Michael Becich - becichmj@msx.upmc.edu;
John Gilbertson* - gilbertsonjr@msx.upmc.edu
* Corresponding author

Published: 13 May 2005

BMC Cancer 2005, 5:45 doi:10.1186/1471-2407-5-45

Received: 08 November 2004
Accepted: 13 May 2005

This article is available from: http://www.biomedcentral.com/1471-2407/5/45

Abstract

Background: Typical high throughput microarrays experiments compare gene expression across two specimen classes – an experimental class and baseline (or comparison) class. The choice of specimen classes is a major factor in the differential gene expression patterns revealed by these experiments. In most studies of prostate cancer, histologically malignant tissue is chosen as the experimental class while normal appearing prostate tissue adjacent to the tumor (adjacent normal) is chosen as the baseline against which comparison is made. However, normal appearing prostate tissue from tumor free organ donors represents an alterative source of baseline tissue for differential expression studies.

Methods: To examine the effect of using donor normal tissue as opposed to adjacent normal tissue as a baseline for prostate cancer expression studies, we compared, using oligonucleotide microarrays, the expression profiles of primary prostate cancer (tumor), adjacent normal tissue and normal tissue from tumor free donors.

Results: Statistical analysis using Significance Analysis of Microarrays (SAM) demonstrates the presence of unique gene expression profiles for each of these specimen classes. The tumor v donor expression profile was more extensive that the tumor v adjacent normal profile. The differentially expressed gene lists from tumor v donor, tumor v adjacent normal and adjacent normal v donor comparisons were examined to identify regulated genes. When donors were used as the baseline, similar genes are highly regulated in both tumor and adjacent normal tissue. Significantly, both tumor and adjacent normal tissue exhibit significant up regulation of proliferation related genes including transcription factors, signal transducers and growth regulators compared to donor tissue. These genes were not picked up in a direct comparison of tumor and adjacent normal tissues.

Conclusions: The up-regulation of these gene types in both tissue types is an unexpected finding and suggests that normal appearing prostate tissue can undergo genetic changes in response to or in expectation of morphologic cancer. A possible field effect surrounding prostate cancers and the implications of these findings for characterizing gene expression changes in prostate tumors are discussed.

Background

Prostate cancer is the most common cancer in men resulting in over 30,000 deaths annually [1]. Early detection and treatment has the potential to markedly reduce the morbidity and mortality associated with the disease. While elevated Prostate Specific Antigen (PSA) [2] is the best available indicator of men with cancer [3], its diagnostic utility is limited due to elevated PSA levels in other non-malignant prostate conditions, varying levels in advanced disease and poor correlation between PSA levels and extent of disease. Furthermore, the variable course of prostate cancer – many patients will not die of the disease – means that radical therapy for all early cases would result in over treatment of significant number of patients. High throughput genomic technologies, by simultaneously interrogating the expression levels of thousands of genes, offers the potential to identify new biomarkers for early detection, prognosis, targets for therapy and for reclassification of prostate tumors. Using expression microarrays, a number of studies have characterized expression profiles for prostate cancer and other tumors. In some cases, correlations between tumor expression signatures, clinical parameters and outcome [4-12] have been identified. While potentially powerful, such studies can be significantly impacted by the choice of baseline or "normal" tissue used to detect tumor related expression changes. Most prostate expression studies to date have utilized normal appearing tissue adjacent to tumor as the tissue for comparison. However, a variety of methods such as chromosomal analysis [13], SAGE [14] and ploidy analysis [15,16] have shown molecular abnormalities in normal appearing prostate adjacent to tumor. Even the term "normal appearing" prostate tissue adjacent to tumor may be misleading, as morphologic researchers using quantitative imaging analysis [17-19], have identified morphologic changes in the epithelial nuclei and blood vessels architecture in prostate tissue adjacent to tumor that are not routinely commented upon by pathologists. This suggests that, in some cases, tissues adjacent to cancer, although appearing morphologically normal by traditional microscopic examination, may contain genetic changes associated with the genesis of or reaction to cancer. Therefore, the use of adjacent normal as the baseline tissue for comparative gene expression studies may mask tumor related molecular changes preceding the appearance of histological tumor. More recently, a microarray study from our institution [12], using adjacent normal and tumor samples, describes a potential field effect around prostate cancers and regulation of selected genes in both adjacent normals and tumors. Using the same microarray data set for our analysis, we have compared the gene expression profiles of prostate cancer, normal appearing prostate tissue adjacent to tumor, and normal appearing prostate tissue from cancer free tissue donors with the aim of identifying the optimal baseline tissue for expression studies and the gene expression changes between the three specimen types.

Methods

Clinical profile of cases

The 60 tumor samples used in this study consisted of 2, 13, 27, 6 and 12 cases of primary prostatic adenocarcinoma of Gleason grade 5, 6, 7, 8 and 9 respectively. There were 4, 20, 23 and 13 cases spanning the age groups 40–49, 50–59, 60–69 and 70–79 respectively. Of the cases, 36 were stage T3 or higher with 2, 22, 23, 11 and 2 cases of stage T2a, T2b, T3a, T3b and 4 respectively.

The 63 adjacent normal samples consisted of 2, 11, 29, 8, and 13 cases of Gleason grade 5, 6, 7, 8 and 9 respectively. There were 4, 21, 25 and 12 cases spanning the age groups 40–49, 50–59, 60–69 and 70–79 respectively. There were 2, 21, 26, 12 and 2 cases of Stage T2a, T2b, T3a, T3b and T4 respectively.

Of the donors, 11 are under and 7 are over the age of 40.

Samples and sample procurement

The tumor and adjacent normal tissue samples were acquired from the University of Pittsburgh Medical Center under stringent Institutional Review Board guidelines with appropriate informed consent. Specimens were received directly from the operating room. Samples (>500 mg) were excised and snap frozen in liquid nitrogen within 30 min of excision and stored at -80°C in the University of Pittsburgh Pathology Tissue Bank until extraction of RNA. All samples were submitted for pathology evaluation. In every case, the tissue was excised from the junction between the ejaculatory duct and the prostatic urethra in the transition zone of the prostate. In particular, adjacent normal tissue was excised away from the cancer lesion macroscopically, and their histological diagnosis was confirmed microscopically.

Donor tissue specimens were received through a collaborative arrangement with the Center for Organ Recovery and Education (CORE), the local organ procurement agency. The arrangement allows the University of Pittsburgh Pathology Tissue Bank to acquire normal prostates and associated serum/plasma specimens from healthy individuals who have donated their organs for transplant. There is extensive collaborative support from CORE. The donor prostatectomies harvested from brain dead, perfused donors and are bathed in Ringer's Lactate solution and transported on wet ice. These donor prostates are transported and handled with the harvested "transplant" organs. This significantly reduces transit time and minimizes the degradation of RNA. The processing methodologies used consist of snap freezing tissues in bulk, freezing in OCT and processing the tissues for routine histology (paraffin embedded tissues. For microarray analysis, the donor samples were excised from the same zone as the tumor and adjacent normal samples.

cRNA preparation and Affymetrix chip hybridization

cRNA was prepared and hybridized to Affymetrix oligonucleotide arrays as previously described [12]

Statistical analysis

We analyzed prostate tissue samples from 18 donors and 63 prostate cancer patients. From the prostate cancer patients we took samples from the histologic tumor as well as normal appearing tissue adjacent to the tumor. High quality RNA and chip data were obtained from 60 cancer and 63 adjacent to tumor samples. In total 141 samples were run against the Affymetrix U95A chip and analyzed. The raw scanned array images were first processed through the Affymetrix Microarray Analysis Suite 5.0 (Affymetrix Corporation, Santa Clara, Ca) to generate probe cel intensity (*.cel) files. The *.cel files were then analyzed using both MAS 5.0 and dChip software from Harvard University [20], to generate gene expression signal values for each probe set. Data normalization to remove variation in overall chip intensities was perfumed by global scaling to a chip mean target intensity of 200 (MAS 5.0) or by the rank-invariant method (dCHIP). The MAS 5.0 with global scaling data and dChip with rank-invariant normalization data gave similar results in the subsequent analysis. Therefore, in the interests of clarity, we will focus on the MAS 5.0 results in the remainder this paper.

In the next phase of analysis, the donors, adjacent normals and tumors were compared for differences in gene expression by using signal values for all 12625 probe sets for each sample. For statistical analysis, we used the Significance Analysis of Microarray (SAM) software package from Stanford University [21]. This method was chosen over conventional statistical tests because of its acceptance in the microarray community, its general simplicity and its ability to provide an estimate of the false discovery rate (the ratio of false positives to total positives). The false discovery rate is particularly important when comparing the expression of thousands of genes simultaneously. For example, when using the Student's t test at a P value of 0.05 to examine a population of 10,000 genes, one would expect 500 false positives. If there were in fact 100 true positives, the false positive rate would be an unacceptably high 0.83 (500/600).

Briefly, SAM calculates a value for each probe set on the array. This value represents the observed difference in mean expression levels between the specimen classes being compared (i.e. tumor and donor) divided by the variance in the data and a fudge factor (see the original paper for details [21]. The resulting value is called the "observed d value". To determine the significance of this value, SAM estimates the "expected" d value if there were no difference between the specimen classes. This is done by permutating (randomly changing) the class labels without changing the data and recalculating the SAM value for each probe set. After thousands of permutations, the result estimates the value that would be obtained if the difference in gene expression were due to chance alone. This is the "expected d value".

The significance of the observed differential gene expression can be estimated by comparing the observed and expected d values. A user defined threshold or "delta" (observed d value – expected d value) can be adjusted to select only those genes observed d value exceeds (for up regulated genes) or is lower (for down regulated genes) than delta. The greater the "delta", the greater the stringency of the result and lower the false discovery rate. For each delta value, the SAM output consists of a gene (probe set) list and an associated false discovery rate. The false discovery rate is estimated from the distribution of expected and observed d values. Probe sets are ordered on the basis of observed d value metric, probe sets with high (or low) values represent genes with relatively high differential expression. The "SAM Plots" are also very useful in visualizing differences in overall differential gene expression between specimen classes.

Results

Expression analysis of tumor, adjacent normal and donor tissue

The prostate tumors analyzed in this study consisted of Gleason grades 5, 6, 7, 8, 9 and patients spanned the ages of 40 through 79. The goal of our research was to examine the differential gene expression patterns observed when comparing our three specimen classes: tumor versus adjacent normal, tumor versus donor and adjacent normal versus donor. The comparison was made at three points in the analytical process: 1) after normalization to remove variation in overall chip intensity 2) after statistical analysis of the data and, 3) after examining the differentially expressed gene lists.

To examine differences in normalized gene expression between tumors, adjacent normals and donors the mean MAS 5.0 and dChip generated signal of each probe set for each specimen class (60 tumors, 63 adjacent normals and 19 donors), was calculated and plotted on a series of scatter plots (Fig 1).

Figure 1 shows the scatter plots and Pearson correlations of the normalized expression data analyzed using both MAS 5.0 and dChip as described above (vide supra, Methods). Data scatter is maximum in the tumor versus donor

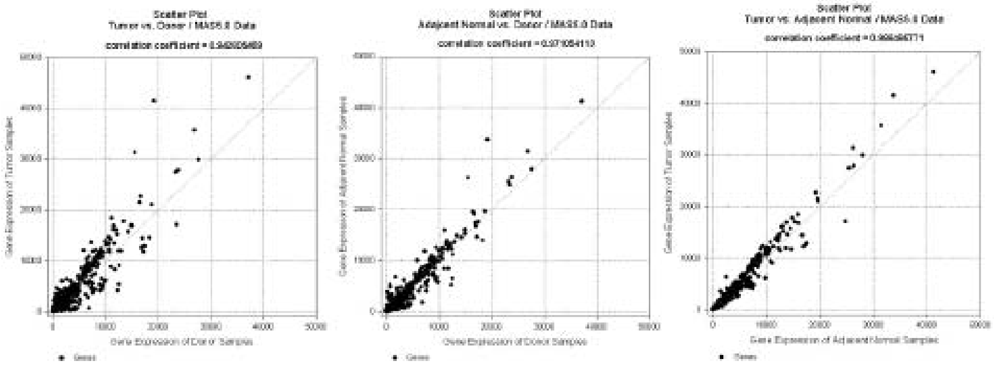


Figure 1

Differential gene expression analysis of donor, tumor and adjacent normal protate cancer samples. Scatter plot of MAS 5.0 derived tumor vs donor, adjacent normal v donor and tumor v adjacent normal samples. For each probe set, the mean MAS 5.0 expression values of all the samples in each specimen group was calculated. Scatter plot were constructed using the mean values for each specimen group.

comparison, intermediate in adjacent normal versus donor and minimal in tumor versus adjacent normal. These findings are suggestive of more differential gene expression in tumor versus donor than tumor versus adjacent normal. In other words, donor normal tissue and adjacent normal tissue do not show the same degree of differential gene expression when paired with tumor tissue. Another striking result apparent in Figure 1 is the close correlation and limited scatter of the tumor versus adjacent plot, even at low levels of signal.

Tumor and adjacent normal specimens came from the same population of patients while donor specimens were received from a different set of individuals. To examine potential patient specific expression effects, the 60 tumors and 63 adjacent normal cases were randomly segmented into two groups, one group provided just tumor data and the other just adjacent normal data and a scatter plot of expression was generated. Since the segmentation of 63 cases can be performed in many different ways (permutations), the scatter analysis was performed 1000 times and the correlation between the sample groups determined by obtaining the mean correlation coefficient of the 1000 permutations. (Figure 2). In this analysis, the close correlation in expression between tumor and adjacent normal specimens persisted even when tumor and adjacent normal samples were taken from different patients.

To determine the statistical significance of the observed differential expression between the three specimen groups, SAM analysis was performed. From each comparison (tumor v adjacent normal, tumor v donor and adjacent normal v donor), a SAM plot was generated and the plots for the three comparisons were overlaid (Fig 3). The diagonal line in Figure 3 represents no differential expression (identical observed and expected d values, for further details see Materials and Methods) with points displaced from the diagonal representing differential expression. Figure 3 shows that each of the comparisons yields a distinct expression profile with donor v tumor exhibiting more differential expression than adjacent normal v tumor or donor v adjacent normal.

To further characterize the expression profiles from these comparisons, differentially expressed gene lists were created from each comparison by selecting genes whose d values (for details, see Materials and Methods) exceed a given threshold. False discovery rates (false positives/ total number of genes in gene list) is no greater than 2.5% at the deltas chosen for this analysis (Table 1).

At a delta of 2.0, when tumor expression is compared to donor expression (Table 1), 474 differentially regulated genes can be detected. At the same delta, when tumor expression is compared to adjacent normals, only 92

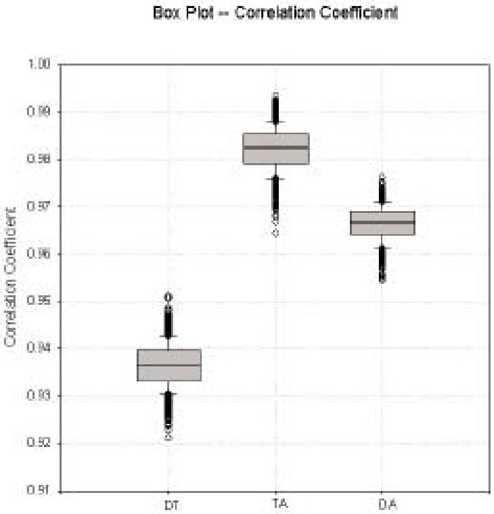


Figure 2
Regression analysis of permuted donors, adjacent normal and tumor samples. The 60 tumors and 63 adjacent normal tissues were segmented so that tumors and adjacent normal samples in each comparison were selected from different patients. The resulting tumor and adjacent normal samples were then subjected to regression analysis. Donor v tumor, donor v adjacent normal and adjacent normal comparisons were performed. Since the segmentation can be performed in many different ways, the analysis was performed 1000 times. The mean correlation coeifficient and standard deviation from each of these comparisons were plotted as box plots.

genes are differentially regulated between these two tissues. Furthermore, at this delta, comparison of tumor expression with adjacent normals does not yield any genes up-regulated in tumors whereas the comparison with donors demonstrates up-regulation of 121 genes. Similarly at other deltas, approximately three times more differentially regulated genes can be detected when tumors are compared to donors than to adjacent normals.

As was discussed above, tumors and adjacent normal tissues are obtained from the same patients and donor tissues from a different sample population. Therefore the larger gene expression differences between tumors and donors may represent underlying patient specific (genetic, demographic or handling) differences in patient (tumor and adjacent normal) and donor prostates rather than intrinsic differences between tumor, adjacent normal and donor normal tissues. It is significant however, that SAM analysis indicates that adjacent normals exhibit far less differential regulation than tumors when both are compared to donors. At all deltas (Table 1), tumors v donors exhibit greater differential expression than adjacent normals v donor implying that tumors and adjacent normals are not identical in gene expression. Therefore, tumor specific, and not patient specific, expression changes can indeed by detected by comparing tumors to donor prostates. Significantly, these results establish the presence of unique gene expression profiles for prostate tissue from donors, adjacent normals and tumors (see Fig 3) with tumors differing more from donors than from adjacent normals.

A potential limitation of our data is that donors span the ages of 5 to 60 and all tumor patients are older than 40. Therefore, the differential gene expression between donors and patients may be due to age specific differences in their prostates. To examine this, we segmented the donors into different age groups and compared only the 40 to 60 year old donors with tumors of the same age group. Although the number of cases in the study were small, the expression pattern observed in this age matched analysis is identical (data not shown) to that when all donors are included suggesting that potential age related differences in donor prostates do not contribute to the results of the donor v tumor analysis.

GO annotation of differential gene expression

We examined the gene lists produced by SAM analysis of tumor, adjacent normal and donor tissue with two objectives: 1) to identify and functionally annotate some of the genes that contribute to the unique expression signatures of these tissues and 2) to determine whether adjacent normals or donors are the more appropriate baseline tissue for detecting differentially expressed genes in tumors. Functional annotation and comparison of the gene lists was performed using Gene Ontology terms [22], for biological processes and Affymetrix's Gene Ontology Mining Tool http://www.affymetrix.com.

When donors are used as the baseline for comparison, tumors exhibit up-regulation of proliferation related genes including transcription factors, signal transducers and growth regulators (see additional file 1). This list includes putative oncogenes, signal transducers and growth regulators. Some of the most up-regulated genes are *v-fos, jun B, jun D, c-src tyrosine kinase, FGF receptor activating protein, immediate early protein* and *early growth response 1*. The most down-regulated genes in tumors include those involved in immune response and signal transduction. Some of the genes in this list are the *interferon induced transmembrane proteins, Duffy blood group antigen* and *tumor necrosis factor*. In contrast, when

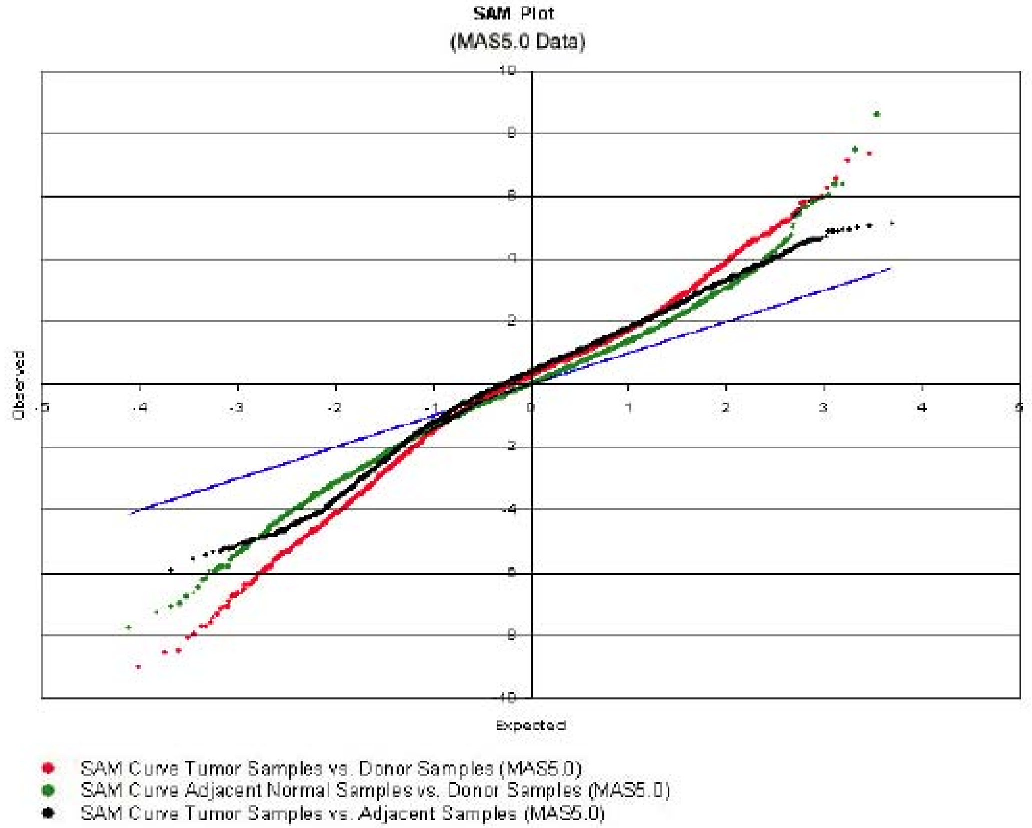


Figure 3

Overlayed SAM plots (for details, see Materials and Methods) from the donor v tumor, donor v adjacent normal and tumor v adjacent normal analyses. Each of the SAM plots was overlayed to direct comparison of the plots. The diagonal line represents no differential gene expression where the observed d value equals the expected d value after 1000 permutations of the class labels. Genes that are differentially expressed are displaced from the diagonal (greater than 0 for up regulation and less than 0 for down regulation). Genes that are more differentially expressed are more displaced from the diagonal than those that are closer to the diagonal. For each of the comparisons, a plot is generated from the d values of the 12625 probe sets in the two specimen groups. Red = donor v tumor plot; green = adjacent normal v tumor plot; black = adjacent normal v donor plot.

adjacent normal tissue is used as the baseline for comparison, tumor tissue exhibits far fewer differentially expressed genes and the genes themselves are less compelling. The list up regulated genes is dominated by ribosomal proteins and metabolic enzymes, while the down-regulated list includes muscle related genes such as *tropomysin*, *actin* and *actinin*.

When expression in adjacent normal is compared to donors, an up-regulation pattern remarkably similar to tumors is seen (additional file 1). Adjacent normals also

Table 1: Differential gene expression in the tumor v donor, tumor v adjacent normal and adjacent normal v donor comparisons. The number of genes identified as differentially regulated at each delta (observed d value – expected d value; for details, see Materials and Methods) are shown. Also, shown are the number of up and down regulated genes at each delta. For each of these deltas, the false discovery rate was no greater than 2.5%.

	Delta 3.0	Delta 2.0	Delta 1.56
Tumor v Donor	65	474	928
Up Regulated	10	121	305
Down Regulated	55	353	623
Tumor v Adjacent Normal	0	92	382
Up Regulated		0	58
Down Regulated		92	324
Adjacent Normal v Donor	12	86	254
Up Regulated	5	25	63
Down Regulated	7	61	191

exhibit up-regulation of putative oncogenes, signal transducers and growth regulators with an almost 70% overlap of the 50 and 100 most up-regulated genes in tumors and adjacent normals, respectively. Similarly there is almost 60% overlap between the most down-regulated genes in tumors and adjacent normals that includes genes involved in immune response.

The biological processes regulated in tumors and adjacent normals were also studied using Affymetrix's Gene Ontology Mining tool. The up regulated gene lists obtained at a SAM delta of 2.0 (Table 1) were uploaded to the tool and the resulting annotations examined. Comparison of tumor gene expression to donor expression reveals up-regulation of genes involved in a number of biological processes (Figure 4a). Amongst these are genes involved in apoptosis, cell cycle, cell proliferation, immune response, protein phosphorylation, protein biosynthesis and transcription. A subset of these including genes involved in immune response and transcription are also up-regulated in adjacent normals (Figure 4b). In contrast when tumor expression is compared to adjacent normals, up-regulation of majority of these processes, except protein metabolism, is not detected (Figure 4c).

Two important conclusions can be derived from the gene annotations, 1) though there are large number of genes regulated in tumors, there is a relatively small subset of genes including oncogenes and signal transducers that are highly regulated in both adjacent normal and tumor tissues and 2) regulation of a number of potentially important biological processes in tumors can be detected from using donors as the baseline tissues. The common regulation of oncogenes, signal transducers and immune response genes in adjacent normals is a striking result in that it suggests that adjacent normal tissue although appearing morphologically normal, undergo gene expression changes that may be important in tumorigenesis or as a reaction to tumor. Since these genes are regulated in both tumor and adjacent normal, they are not picked up on a direct comparison of the two tissues. While it is possible that donors are different from both adjacent normals and tumors due to processing artifacts – the tumor and adjacent tissues were taken at surgery and donors at harvesting – it is unlikely that the large differences seen in donor v tumor are all due to processing differences. This issue is examined further in the discussion section.

The up regulation of proliferation markers in both adjacent normals and tumors coupled with the result that more differential regulation is detected when tumors are compared to donors than to adjacent normals suggests that donor prostates may be the more appropriate tissue for expression studies. Regulation of critical of biological processes and pathways may remain undetected if tissue adjacent to tumors is used for comparison.

Discussion

There is a growing interest in the use of high throughput microarray analysis for the molecular reclassification of diseases. This interest appears to be well founded, as many groups have reported consistent patterns of gene expression associated with pathologic phenotypes, clinical behaviors and outcomes [4-11]. In the area of prostate cancer numerous groups [23-29] have all reported significant differential gene expression between histologic tumor specimens and normal appearing prostate tissue from patients with tumor present elsewhere in the prostate. Recently, a group from our institution reported a 70 gene signature that may predict aggressiveness of prostate cancer [12]. Comparison of the gene lists from published data sets with the results of our tumor versus adjacent normal analysis is complicated by the heterogeneity in samples, analysis platforms and analysis methods. Nevertheless, our study is qualitatively similar to other studies in the expression profile of tumors compared to adjacent normal tissue. A number of genes including *hepsin, myc, fatty acid synthase SPARC1* and *EBNA-2* coactivator show similar expression patterns across multiple prostate cancer studies [30], and are also regulated in our study.

Our donors did not have prostate cancer or prostatic intraepithelial neoplasia (PIN) identified in their prostate and as such are good candidates for "true normals". Differential expression was much greater between tumor and donor tissue than between tumor and adjacent normal.

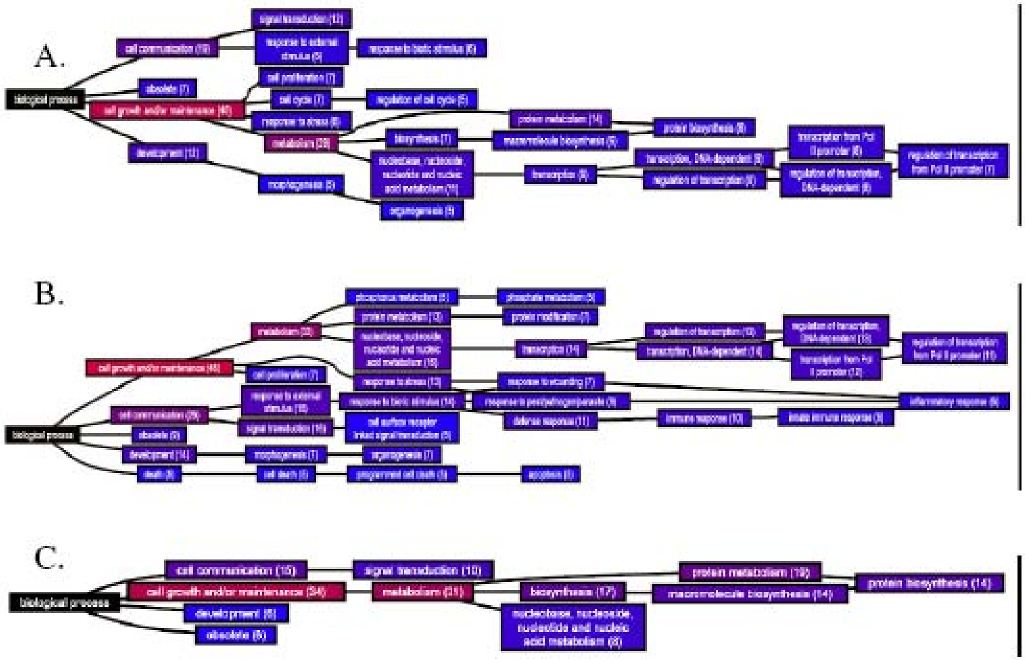

A.
biological process
cell communication (13)
obsolete (7)
cell growth and/or maintenance (40)
development (13)
signal transduction (12)
response to external stimulus (9)
response to biotic stimulus (6)
cell proliferation (7)
cell cycle (7)
regulation of cell cycle (5)
response to stress (8)
metabolism (28)
biosynthesis (7)
nucleobase, nucleoside, nucleotide and nucleic acid metabolism (15)
morphogenesis (8)
organogenesis (5)
protein metabolism (14)
macromolecule biosynthesis (6)
protein biosynthesis (6)
transcription (9)
transcription, DNA-dependent (9)
regulation of transcription (8)
transcription from Pol II promoter (8)
regulation of transcription, DNA-dependent (8)
regulation of transcription from Pol II promoter (7)
B.
biological process
cell growth and/or maintenance (46)
metabolism (33)
cell communication (29)
obsolete (5)
development (14)
death (8)
phosphorus metabolism (5)
protein metabolism (13)
nucleobase, nucleoside, nucleotide and nucleic acid metabolism (16)
cell proliferation (7)
response to stress (13)
response to external stimulus (18)
response to biotic stimulus (14)
signal transduction (16)
cell surface receptor linked signal transduction (5)
morphogenesis (7)
organogenesis (7)
cell death (8)
programmed cell death (8)
apoptosis (8)
phosphate metabolism (5)
protein modification (7)
transcription (14)
response to wounding (7)
response to pest/pathogen/parasite (9)
defense response (11)
regulation of transcription (13)
transcription, DNA-dependent (14)
immune response (10)
regulation of transcription, DNA-dependent (13)
transcription from Pol II promoter (12)
innate immune response (5)
regulation of transcription from Pol II promoter (11)
inflammatory response (6)
C.
biological process
cell communication (15)
cell growth and/or maintenance (34)
development (8)
obsolete (6)
signal transduction (10)
metabolism (31)
protein metabolism (19)
biosynthesis (17)
macromolecule biosynthesis (14)
protein biosynthesis (14)
nucleobase, nucleoside, nucleotide and nucleic acid metabolism (6)

Figure 4

Gene Ontology annotation of differentially expressed gene lists. The fifty most upregulated genes from the donor v tumor, adjacent normal v tumor and tumor v adjacent normal comparisons were uploaded to Affymetrix's Gene Ontology Mining Tool, *a*, donor v tumor; *b*, donor v adjacent normal; *c* adjacent normal v tumor; The annotations is presented as a hierarchy of terms, from general to most specific terms (from left to right). The numbers in parenthesis indicate the number of genes that are annotated with the term. In all of the analysis, annotation of all the submitted probe sets is not achieved. Typically, annotation exists for approximately 60% of the probe sets.

The fact that tumor and adjacent specimens come from the same patients could possibly explain this difference but this was ruled out by our analysis. Another possibility is that tissue handing and processing differences could account for some or all of the differential expression seen when donor tissue is use as a baseline. In fact, data in the literature does suggest that tissue processing could effect the expression of genes such as *fos*, *jun* and *egr* in prostate tissue [29]. However, the same literature indicates that the effect warm ischemic time is limited to specific genes and in general, involves less than 1% of the regulated genes [29]. Our studies emphasize the need for documentation and quality of all experimental processing steps, from sample acquisition to sample hybridization, in order to completely characterize gene expression differences between prostate donors, tumors and normal tissue adjacent to tumors.

In our experiment, tumor and adjacent normal specimens where taken from the same prostates and handled the same way. If differences in patient and donor tissue handling was the major issue driving differential expression in the tumor v donor and adjacent normal v donor comparisons, one would expect tumor v donor and adjacent normal to result in very similar expression profiles. However, we have shown that tumor v donor exhibits far greater differential expression than adjacent normal v donor (see Results). Furthermore, the differentially expressed genes seen in both tumor and adjacent normal include proto-oncogene and transcription factors that one might rationally expect to see in expectation of or in response to a local tumor. Therefore, while the possibility that some expression changes are due to differences in tissue handling cannot be formally ruled out, it is unlikely that the large and specific differences we observe in tumor v donor, tumor v adjacent normal and adjacent normal v donor are entirely due to processing differences. Clearly additional studies, including examination of patient process specimens that do not host prostate cancer (such as cystoprostatectomy for bladder cancer or prostates removed for benign hypertrophy) to examine this process further.

The most important finding from our analysis is the potential importance of the donor specimens and the possibility that a field effect exists around prostate tumors, resulting in significant molecular changes in histologically normal appearing tissue adjacent to prostate cancer. Significantly, evidence for such malignancy associated changes have been presented in other organs such as the cervix, bladder and breast [31-33]

Furthermore, a variety of methods such as chromosomal analysis [13], SAGE [14], ploidy analysis [15,16] have shown molecular abnormalities in normal appearing prostate adjacent to tumor. Image analysis has also been employed to identify consistent changes in "normal appearing" prostate tissue adjacent to tumor [17,18]. In one study cases of prostatic adenocarcinoma was consistently detected by examining histologically normal tissue using high-resolution image cytometry [18], and in another, combined highly sensitive and discriminating Fourier transform-infrared spectroscopy with statistical analysis was used to detect damaged DNA in normal appearing prostate tissue adjacent to cancer [34].

In expression analysis, while most published prostate studies have used adjacent normals as the baseline tissue, Dhansekaran [23], used both commercially available pooled donor normal tissue and adjacent normal tissue and noted differences in expression profile between the two specimen types. Genes that were differentially expressed in adjacent normals when compared to the pooled donor normals included signal transducers and transcription factors; and expression of these genes in adjacent normals was attributed to a field effect around tumors. Similarly, Yu [12] have noted dysregulation of selected genes in both adjacent normals and tumors when compared to donors. Prakash [27], found that gene expression in asymptomatic benign prostatic hyperplasia adjacent to tumors was different from asymptomatic BPH or symptomatic BPH not associated with tumors. The unique expression signature of BPH next to tumors included fos, jun, immediate early genes and this list was remarkably similar to the most up-regulated genes in the adjacent normals tissue in our study (see adjacent normal v donor, additional file 1).

Finally, within archives of the University of Pittsburgh Pathology Tissue Bank, there was a donor prostate, which was found to harbor prostate cancer. When run on the Affymetrix arrays, the tumor classified with the tumors samples rather than the donor samples. Although this is clearly no more than an anecdotal event, it is an interesting finding.

Though microarray technology represents a major advance and provides a powerful tool for high-throughput expression analysis, the most effective use of this technology requires careful consideration of baseline normal tissue. Our results here emphasize the need for careful examination of what constitutes normal tissue and the importance of future studies to fully characterize normal appearing tissue adjacent to prostate cancers.

Conclusions

Prostate tumor tissue, histologically normal tissue adjacent to tumors and donor normal prostate tissue exhibit unique gene expression profiles with tumor and adjacent normal profiles more similar to each other than to the donors. These results suggest that normal appearing tissue around prostate tumors may also be undergoing tumor related changes and that careful characterization of these different tissues is necessary to understand molecular changes in leading up to prostate cancer.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

URC was involved in the data analysis and interpretation and manuscript writing and revision. RD, as director of the Tissue Bank was involved in obtaining IRB approval and providing deidentified tissue for this research study. CM was involved in data analysis and interpretation. GM and MB as principal investigator and co-investigators of the grant titled "the Molecular Reclassification of Cancer" were responsible for design of the study, providing intellectual input and final approval of the version submitted for review. JG was responsible for providing intellectual direction, guidance for the analysis team, review and final approval of the manuscript.

Additional material

Additional File 1

The fifty most up regulated and down regulated genes from the tumor v donor, tumor v adjacent normal and adjacent normal v donor comparison. The Affymetrix probe set id, gene names and assignment of biological process for each gene are shown. The biological process annotation includes information from Affymetrix, gene ontology and literature references.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2407-5-45-S1.xls]

Acknowledgements

This work was funded by 1 U01 CA86735-01 (to MB)"A Shared Resource for the Molecular Classification of Prostate Cancer Comprehensive Prostate Cancer Tissue Resource (CPCTR) Consortium, National Cancer Institute, 5 U01 CA88110-02 (to GM, MB is Co-Principal Investigator)

"Molecular Reclassification of Prostate Cancer" Director's Challenge for the Molecular Classification of Cancer Consortium National Cancer Institute and 2000-98 (to JG) DOD National Medical Technology Test Bed An Advanced Tissue Banking Information System

References

1. **Cancer facts and figures.** 2003.
2. Sokoll LJ, Chan DW: **Prostate-specific antigen.** *Urology Clinics North America* 1997, 24:253-259.
3. Polasckik TJ, Oesterling JE, Partin AW: **Prostate specific antigen. Its discovery and biochemical characteristics.** *Journal of Urology* 1999, 162:293-306.
4. Takahashi M, Rhodes DR, Furge KA, Kanayama H, Kagawa S, Haab BB: **Gene expression profiling of clear renal cell carcinoma: gene identification and prognostic classification.** *PNAS* 2001, 98:9754-9759.
5. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, 415(6871):530-536.
6. Pomeroy S, Tamayo P, Gaasenbeek M, Sturla L, Mangelo M, McLaughlin ME, Kim JYH, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR: **Prediction of central nervous system embryonal tumoru outcome based on gene expression.** *Nature* 2002, 415:436-442.
7. Geisler SJH, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botsten D, Lonning PE, Borresen-Dale A-L: **Gene expression patterns of breast carcinomas distinguish tumor subcalsses with clinical implications.** *PNAS* 2001, 98:10869-10874.
8. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S: **Gene-experssion profiles predict survival of patients with lung adenocarcinoma.** *Nature Medicine* 2002, 8(8):816-824.
9. Golub TR, Slomin DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller HLM, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: ***Molecular classification of cancer: class discovery and class prediciton by gene expression monitoring.*** *Science* 1999, 286:531-537.
10. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *PNAS* 2001, 98:13790-13795.
11. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda Massimo, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures.** *PNAS* 2001, 98(26):15149-15154.
12. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C, Thomas R, Dhir R, Finkelstein S, Michalopoulos G, Becich M, Luo JH: **Gene Expression Alterations in Prostate Cancer Predicting Tumor Aggression and Preceding Development of Malignancy.** *Journal of Clinical Oncology* 2004, 2:2790-2799.
13. Sandberg AA: **Chrosomal abnormalities and related events in prostate cancer.** *Human Pathology* 1992, 23(4):368-380.
14. Waghray A, Schober M, Feroze F, Yao F, Virgin J, Chen YQ: **Identification of differentially expressed genes by serial analysis of gene epxression in human prostate cancer.** *Cancer Research* 2001, 61:4283-4286.
15. Bostwick DJ: **Prospective origins of prostate carcinoma. Prostatic intraepithelial neoplasia and atypical adenomatous hyperplasia.** *Cancer Reserach* 1996, 78:330-336.
16. Bostwick DJ, Shan A, Qian J, Darson M, Maihle NJ, Jenkins RB, Cheng L: **Independent origin of multiple foci of prostatic intraepithelial neoplasia.** *Cancer* 2000, 83(9):1995-2002.
17. Montironi R, Diaminiti L, Santinelli A, Magi GC, Scarpelli M, Giannulis I, Mangili F: **Subtle changes in benign tissue adjacent to prostate neoplasia detected with a bayesian belief network.** *Journal of Pathology* 1997, 182:442-449.
18. Bartels PH, Montironi R, Duval da Silva V, Hamilton PW: **Tissue architecture analysis of prostate cancer and its precursors: an innovative approach to computerized histometry.** *European Urology* 1999, 35:484-491.
19. Mairinger TG, Mikuz G, Gschwendtner : **Nuclear chromatin texture analysis of nonmalignant tissue can detect adjacent prostatic adenocarcinoma.** *The Prostate* 1999, 41:12-19.
20. Schadt E, Li C, Blis B, Wong WH: **Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.** *Journal of Cellular Biochemistry Supplement* 2001, 37:120-125.
21. Tusher VGTR, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *PNAS* 2001, 98(9):5116-5121.
22. Ashburner M, Ball C, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The gene ontology consortium.** *Nature Genetics* 2000, 1:25-29.
23. Dhanasekaran SM, Barette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: **Delineation of prognostic biomarkers in prostate cancer.** *Nature* 2001, 412:822-826.
24. Luo J, Duggan D, Chen Y, Sauvageot J, Ewing CM, Bittner ML, Trent JM, Isaacs WB: **Human Prostate Cancer and Benign Prostatic Hyperplasia: Molecular Dissection by Gene Expression Profiling.** *Cancer Reserach* 2001, 61:4683-4688.
25. Luo JH, Yu YP, Cieply K, Lin F, Deflavia P, Dhir R, Finkelstein S, Michalopoulos G, Becich M: **Gene expression analysis of prostate cancers.** *Molecular Carcinogenesis* 2002, 33:25-35.
26. Chetcuti A, Margan S, Mann S, Russell P, Handelsman D, Rogers J, Dong : **Identification of differentialy expressed genes in organ-confined prostate cancer by gene expression array.** 2001, 47:132-140.
27. Prakash K, Pirozzi G, Elashoff M, Munger W, Waga I, Dhir R, Kakehi Y, Getzenberg RH: **Symptomatic and asymptomatic benign prostatic hyperplasia: Molecular differentiation by using microarrays.** *PNAS* 2002, 99(11):7598-7603.
28. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub T, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, 1:203-209.
29. Dash A, Maine IP, Varambally S, Shen R, Chinnaiyan AM, Rubin MA: **Changes in differential gene expression because of warm ischemia time of radical prostatectomy specimens.** *Am J Pathol* 2002, 161:1743-1748.
30. Rhodes DR, Barett TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expres-**

sion profiles reveals pathway dysregulation in prostate cancer. *Cancer Research* 2002, 62:4427-4433.

31. MacAulay C, Lam S, Payne PW, LeRiche JC, Palcic B: **Malignancy-associated changes in bronchial epithelial cells in biopsy specimens.** *Anal Quant Cytol Histol* 1995, 17:55-61.
32. Montag AG, Bartels P, Dytch HE, Lerma-Puertas E, Michelassi F, Bibbo M.: **Karyometric marker features in tissue adjacent to in situ cervical carcinoma.** *Anal Quant Cytol Histol* 1989, 11:275-280.
33. Montag AG, Bartels PH, Dytch HE, Lerma-Puertas E, Michelassi F, Bibbo M: **Karyometric features in nuclei near colonic adenocarcinoma. Statistical analysis.** *Anal Quant Cytol Histol* 1991, 13:159-167.
34. Malins DC, Johnson PM, Barker EA, Polissar NL, Wheeler TM, Anderson KM: **Cancer-related changes in prostate DNA as men age and early identification of metastasis in primary prostate tumors.** *PNAS* 2003, 100(9):5401-5406.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2407/5/45/prepub

APPENDIX B

Decision tree learning-based characterization of the global effects of cocaine abuse on gene expression in the rat brain

Decision tree learning-based characterization of the global effects of cocaine abuse on gene expression in the rat brain

Changqing Ma **Vanathi Gopalakrishnan** *

Department of Pathology
University of Pittsburgh
Pittsburgh PA 15260
Tel. (412) 383-7625
Fax (412) 647-6676
chmst40+@pitt.edu

Department of Medicine
University of Pittsburgh
Pittsburgh PA 15260
Tel. (412) 648-6677
Fax (412) 647-7190
vanathi@cbmi.upmc.edu

David G. Peters **Robert E. Ferrell**

Department of Human Genetics
University of Pittsburgh
Pittsburgh PA 15260
{dgp@imap.pitt.edu, rferrell@helix.hgen.pitt.edu}

Submitted to the Intelligence System for Molecular Biology (ISMB) 2002 Conference as a student paper. It was accepted and presented as a poster at the conference.

ABSTRACT

Motivation: *This study aims to characterize the global changes in gene expression across the rat brain due to an acute dose of cocaine. Microarray gene expression data were generated from cocaine-treated and untreated tissue samples from five regions of the rat brain. A decision tree learning method was applied to this data to learn plausible models of the interactions among the brain regions.*

Results: *Our approach to normalization and filtering of the original dataset provides a useful methodology for successful application of decision tree learning to this novel gene expression dataset. The popular decision tree learning program C4.5 learned a highly accurate (97.53% average prediction accuracy from cross-validation) and human-understandable model from the normalized and filtered data. The learned model depicted a global change in gene expression among three brain regions in response to an acute dose of cocaine. The learned global pattern was verified independently using a different normalization procedure and visualization. The rule sets were studied carefully and the genes covered by each rule were annotated based on Gene Ontology terms.*

Contact: chmst40@pitt.edu

Supplementary Information: *The normalized dataset and the Gene Ontology annotations of the genes covered by each rule are available at http://www.pitt.edu/~chmst40/ratdata/.*

Key words: Decision Tree Learning, Cocaine Abuse, Descriptive Generalized Models, Gene Expression Analysis, Normalization.

* To whom correspondence should be addressed.

INTRODUCTION

One of the major public health problems in the United States stems from the abuse of psychostimulant drugs, such as cocaine, amphetamine and their derivatives. A critical property of these drugs is that they tend to be addictive, that is, when administered in acute doses, their usage entails repeated usage. Cocaine and similar drugs induce the expression of immediate early genes which activate several networks of biochemical pathways in brain neurons (Hope, 1998; Torres and Horowitz, 1999). These affected neurons locate in different brain regions and belong to different brain systems but synaptically converge on a common set of mesocorticolimbic neurons (Torres and Horowitz, 1999). The complexity of this system has made it difficult to map gene expression to addictive behaviors. Recently developed high-throughput microarray technology, however, allows the expression of thousands of genes to be monitored simultaneously and thus has the potential to enable the study of drug abuse at the genomic scale. This capability may provide a new way to understand the global changes in the gene expression patterns in brain due to drug abuse. Specifically, we can identify patterns of gene expression that correspond to cocaine exposure.

A number of popular methods exist to identify higher order patterns in gene expression data. These methods can be usefully classified as being "supervised" or "unsupervised". Unsupervised methods, such as clustering (Eisen et al., 1998) and self-organizing maps (Tamayo, et al., 1999), seek to identify patterns in the gene expression data without the use of prior knowledge. Such methods are useful in basic data discovery and often find unique and novel groupings in the data but often do not reproduce known groups. Supervised methods, such as the decision tree learning method used in this study, on the other hand, incorporate informative specimen labels and knowledge about the dataset beyond the gene expression data. For example, a supervised method might be told which subjects had a given disease and which had not, and it would use that information to classify gene expression patterns. Supervised methods are likely to find gene expression patterns that correlate with the external labels, in this case the disease or lack of the disease.

Decision tree learning (Quinlan, 1986) is a commonly used technique to derive plausible descriptive models from training examples that can used to classify test examples whose classification is unknown. A primary advantage with respect to clustering methods and other supervised learning methods is that the predictive models obtained from decision tree learning method are human-understandable rules and therefore, enable characterization of general trends within the training dataset. Decision tree learning has been applied in the past to gene expression data (Brown et al, 2000), but with limited success.

In this paper, we describe an approach to normalization and filtering of a novel gene expression dataset that enables the learning of highly accurate decision rules to characterize gene expression obtained from normal tissue as well after treatment with an acute dose of cocaine. The analysis of this initial set of experimental data aims to understand the global effects of cocaine across the brain, using rat as the model animal. The popular decision tree learning program C4.5 was used to learn a highly accurate decision tree, and then generate set of production rules (using the C4.5rules program) that describe a generalized model for discriminating between cocaine-treated and untreated brain. This model describes the global effect of cocaine in the rat brain and implicates the interaction among the regions in rat brain due to cocaine abuse.

METHOD

The Cocaine Dataset

Forty male Sprague Dawley rats, twenty naïve and twenty sensitized with an acute dose of cocaine were used in the experiment. From the naïve and sensitized cohorts, pools of total RNA, from five brain areas: the Amygdala (AMY), Caudate Putamen (CPU), Nucleus Accumbens (NA), Prefrontal Cortex (PFC) and Ventral Tegmental Area (VTA), were used as the substrates for cDNA synthesis (see Figure 1 for spatial distribution of the brain regions). Region specific tissue from twenty animals was required to procure sufficient high-quality mRNA for the microarray experiments. It was also hoped that pooling tissue from twenty animals would also dampen the possible effect of inter-individual differences in gene expression at baseline or following treatment. We therefore had ten samples for analyzing differences in gene expression using commercially available Rat Genome U34A (RG-U34) array set from Affymetrix, Inc, Santa Clara, CA.

There are totally 8799 probe sets on the RG-U34A array derived from all full-length or annotated genes (~7000) as well as thousands of EST clusters. In this paper, we will use "gene" as the general term referring to the genes and ESTs on the microarray chip. The results of gene expression analysis using the Affymetrix Micro Array Suite 4.0 software from Affymetrix, Inc., Santa Clara, CA, are expressed as several parameters representing both qualitative as well as quantitative information for each gene represented on the arrays. An important quantitative measure of gene expression is represented by the Average Difference parameter. The Average Difference is a relative indicator of the level of expression of a transcript, and is used to determine the changes in expression of a given gene. For each gene, there are ten Average Difference data points corresponding to its relative expression level in cocaine-treated and untreated tissue samples from five brain regions. Thus, there were 87990 data points available for analysis, divided as follows - 8799 genes x 2 conditions x 5 regions.

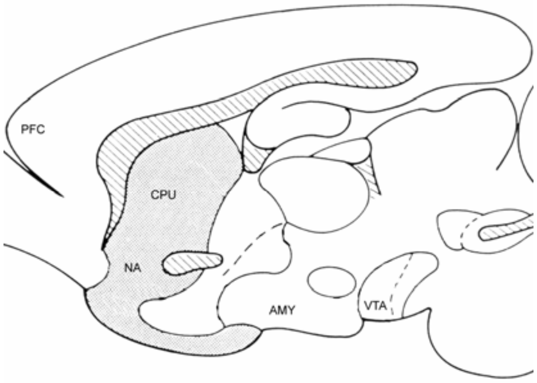


Figure 1. Schematic representation of some of the neurotransmitter systems in the rat brain sagittal section. AMY = Amygdala; CPU = Caudate Putamen; NA = Nucleus Accumbens; PFC = Prefrontal Cortex; VTA = Ventral Tegmental area. The schema is derived from page 55 of the Rat Nervous System volume 1: Forebrain and Midbrain (Paxinos, 1985).

Overall Methodology

Our methodology used for analysis of this dataset is depicted in Figure 2. The Average Difference values from 8799 genes, from ten samples were merged into one data file, referred to as the raw data (Figure 2a, Step 1). Using this data,

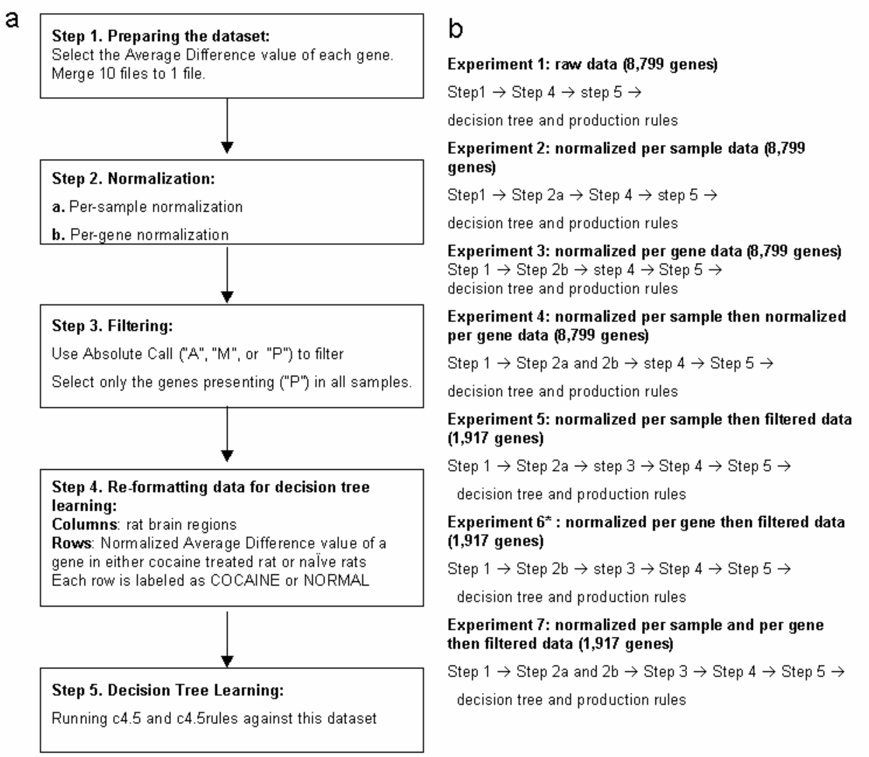

a
Step 1. Preparing the dataset:
Select the Average Difference value of each gene.
Merge 10 files to 1 file.
Step 2. Normalization:
a. Per-sample normalization
b. Per-gene normalization
Step 3. Filtering:
Use Absolute Call ("A", "M", or "P") to filter
Select only the genes presenting ("P") in all samples.
Step 4. Re-formatting data for decision tree learning:
Columns: rat brain regions
Rows: Normalized Average Difference value of a gene in either cocaine treated rat or naÏve rats
Each row is labeled as COCAINE or NORMAL
Step 5. Decision Tree Learning:
Running c4.5 and c4.5rules against this dataset
b
Experiment 1: raw data (8,799 genes)
Step1 → Step 4 → step 5 →
decision tree and production rules
Experiment 2: normalized per sample data (8,799 genes)
Step1 → Step 2a → Step 4 → step 5 →
decision tree and production rules
Experiment 3: normalized per gene data (8,799 genes)
Step 1 → Step 2b → step 4 → Step 5 →
decision tree and production rules
Experiment 4: normalized per sample then normalized per gene data (8,799 genes)
Step 1 → Step 2a and 2b → step 4 → Step 5 →
decision tree and production rules
Experiment 5: normalized per sample then filtered data (1,917 genes)
Step 1 → Step 2a → step 3 → Step 4 → Step 5 →
decision tree and production rules
Experiment 6* : normalized per gene then filtered data (1,917 genes)
Step 1 → Step 2b → step 3 → Step 4 → Step 5 →
decision tree and production rules
Experiment 7: normalized per sample and per gene then filtered data (1,917 genes)
Step 1 → Step 2a and 2b → Step 3 → Step 4 → Step 5 →
decision tree and production rules

Figure.2. Our method for learning models for the cocaine dataset. (a) A flowchart for the possible steps to perform the decision tree learning. A brief explanation is given to each step. (b) The seven experiments we performed by applying different combinations of the steps shown in (a). The * indicates that the sequence of performing normalization and filtering is not critical for deriving the result.

the decision tree learning program C4.5 learned an extremely inaccurate model (Figure 2b, Experiment1). The default settings were used for all runs of C4.5, as changing the parameters did not significantly improve the prediction accuracy of the learned models. The raw data was subsequently normalized and/or filtered prior to decision tree learning (Figure2a, steps 2 and 3; Figure2b, Experiments 2 to 7). Experiment 7 yielded a highly accurate model from normalized and filtered datasets, with a small number of rules within the rule set. The gene expression pattern implicated by this model was observed and validated visually by using GeneSpring 4.1. The genes covered by each rule were annotated on the basis of TIGR Rat Gene Index. The software tools used in this study are listed below:

- C4.5 Release 8 publicly available at http://www.cse.unsw.edu.au/~quinlan/ was used for decision tree learning.
- GeneSpring software version 4.1 from Silicon Genetics, Redwood City, CA, was used to normalize the raw data and perform gene clustering using spearman correlation. The gene tree was used to validate the decision tree learning result.
- TIGR (The Institute for Genomic Research) Rat Gene Index release Version 6.0 is publicly available at http://www.tigr.org/tdb/rgi/index.html. This index provides the source of gene ontology annotation.

Decision Tree Learning

Training examples are represented as the gene expression values for each brain region for each gene (i.e. five values in this dataset), followed by the brain tissues' classification, i.e. normal (untreated) or cocaine (cocaine-treated). Two examples are shown below, where each of the five relative expression values is separated by commas, followed by the target class (NORMAL or COCAINE):

0.949,1.143,0.691,0.242,1.316,COCAINE.
1.006,1.789,0.718,0.783,2.121,NORMAL.

Given such training examples, we apply C4.5 to learn a decision tree classifier that comprises of region-wise tests of expression values that best discriminates examples of cocaine-treated class from the normal brain tissue samples. The decision tree is built incrementally by applying an entropy-based measure called "information gain" to determine which attribute (gene expression in a particular brain region such as Amygdala) is most informative in terms of discriminating between the target classes (i.e. normal vs. cocaine). This most informative attribute is then placed as the first test at the root of the decision tree, with branches labeled according to its values. The training data are then sorted along the branches. For each branch, the next most informative attribute that best discriminates among the subset of training data along that branch, is then chosen as the attribute whose values will be tested. The process continues until there are no more training examples that need to be covered (classified) along each branch. The leaf nodes contain labels of the target class, and represent the classification of the conjunction of features (<attribute, value> pairs) along that unique path from the root attribute of the decision tree. The most general classifier and the smallest decision-tree that does not over-fit the training data is used for prediction (Quinlan, 1993).

Data Normalization

A normalization procedure is first applied to the data (8799 x 10) by using GeneSpring 4.1. To normalize in the context of DNA microarrays means to standardize your data to be able to differentiate real (biological) variations in gene expression levels and variations due to the measurement process. Normalizing also scales the data so that you can compare relative gene expression levels (GeneSpring, 2001). Here, each sample was normalized to itself first, and then each gene was normalized to itself across all the ten data points.

The normalization of each sample to itself (also called per sample normalization or normalized per sample in this paper) intends to remove the differences in amount of exposure between samples, so different samples are comparable to one another. The median of all measurements in a given sample X is set to 1, and all other values scaled accordingly. The formula used is:

$$\frac{\text{(the signal strength of gene A in sample X)}}{\text{(the median of all of the measurements taken in sample X)}}$$

The normalization of each gene to itself (also called per gene normalization or normalized per gene in this paper) accounts for the difference in detection efficiency between spots. It also allows you to compare the relative change in gene expression levels, as well as display these levels in a similar scale on the same graph. The formula used is:

$$\frac{\text{(the signal strength of gene A in sample X)}}{\begin{array}{c}\text{(the median of every measurement taken for gene A}\\\text{throughout all of the samples)}\end{array}}$$

The mean value of all gene expression data, for one given sample or gene, is commonly used in normalization since mean is correlated with standard deviation of that data. However in our dataset, there are no repeat measurements available for each sample. Moreover, the data points for a given gene are obtained from tissues across different rat brain regions. It is not meaningful to determine the standard deviation from this gene expression data, for a given sample or gene, and consequently the mean value of that data is not reliable. Under this condition, the median value of the data is less subject to outliers and therefore more representative of the overall expression level for all the expression data for a given sample or gene. Therefore, the median value was used for normalization instead of mean value in this study.

Filtering

The filtering procedure took advantage of a qualitative feature in the Affymetrix expression data files called the Absolute Call Metric (Affymetrix Inc., 2000). Using the Absolute Call we could eliminate transcripts that were reported to be absent (A) or only marginally present (M) as detected by the technology. The filter was designed to select only genes reported to be present (P) in all 5 brain regions under both cocaine-treated and normal brain conditions. After applying this filter to the original data, we obtained 1917 genes that are present in all ten experiments.

Schadt et al. (2000) have argued that filtering of genes on the basis of Absolute Call can be associated with some risk as genes with "absent" or "marginal" expression may be informative. However, the goal of our study was to detect differences in global expression pattern in cocaine exposed and naïve rats. Therefore, the 3140 genes (36% of the 8799 genes on the chip) that were "absent" in all samples are clearly not informative for our purposes. This left 42% of genes which were called "present" in between 1 and 9 of the samples (22% were expressed in all ten specimens and therefore included by the filter). Because of the large quantitative variation induced by these partially expressed genes, and uncertainty of how to represent "absent" genes in our model, we decided to focus on the 1917 genes that were expressed in all samples.

RESULTS AND DISCUSSION

Prediction Accuracy

Table 1 depicts the dramatic improvement in the prediction accuracy of the classifier due to our method as described in Figure 2. TP and TN refer to the true positive rate (sensitivity) and true negative rate (specificity); FP and FN refer to the false positive and false negative rate. The accuracy of each classifier is calculated as the percentage of training examples that are classified accurately (the number of correct predictions/ the number of examples predicted * 100). As reported in this table, the true positive rate (TP) for cocaine class prediction increased to 98.85% in both Experiment 5 and 7. The accuracy of the classifiers for classifying the training examples increased to 98.36% and 98.88% respectively.

Experiment 5 and 7 yielded the most accurate models. The datasets used in these two experiments were both normalized sample-wise but, in Experiment 7, the data were also

Table.5.1. Comparison of the prediction accuracy of models generated by running C4.5 against different datasets. TP: true positive rate; FP: false positive rate; TN: true negative rate; FN: false negative rate; AC: accuracy. The experiment numbers are identical to those of Figure2b.

Datasets	TP (%)	FP (%)	TN (%)	FN (%)	AC (%)
Experiment 1. raw data (8799 genes, no normalization and no filtering)	45.49	17.14	82.86	54.51	64.18
normalized data (8799 genes in each dataset)					
Experiment 2. normalized per sample data	87.2	24.7	75.3	12.8	81.25
Experiment 3. normalized per gene data	70.2	32.37	67.63	29.8	68.92
Experiment 4. normalized per sample and per gene data	87.16	26.51	73.49	12.84	80.32
normalized and filtered data (1917 genes in each dataset)					
Experiment 5. normalized per sample then filtered data	98.85	2.13	97.86	1.15	98.36
Experiment 6. normalized per gene then filtered data	19.67	6	94	80.33	56.83
Experiment 7. normalized per sample and per gene then filtered data	98.85	1.1	98.9	1.15	98.88

Table 5.2. Results from a ten-fold cross-validation on the training set consisting of 3834 examples (1917 genes x 2 classes). TP: true positive rate; FP: false positive rate; TN: true negative rate; FN: false negative rate; AC: accuracy.

Times	TP (%)	FP (%)	TN (%)	FN (%)	AC (%)
1	98.44	3.12	96.88	1.56	97.66
2	97.92	1.04	98.96	2.08	98.44
3	96.35	2.08	97.92	3.65	97.14
4	93.75	1.04	98.96	6.25	96.35
5	97.92	2.6	97.4	2.08	97.66
6	97.4	1.04	98.96	2.6	98.18
7	98.44	3.65	96.35	1.56	97.4
8	98.44	2.08	97.92	1.56	98.18
9	97.92	3.65	96.35	2.08	97.14
10	97.4	3.12	96.88	2.6	97.14
Average	97.4	2.34	97.66	2.6	97.53

normalized gene-wise prior to the decision tree learning. This indicates that per sample normalization makes a significant difference in prediction accuracy in learned models. The model learned from Experiment 7 was selected as it has the smaller set of rules (17 rules) compared with the model learned from Experiment 5 (37 rules).

Cross-validation

To test the robustness of the selected model, a ten-fold cross-validation is performed, wherein 10 separate runs of C4.5 were made using each time 90% of the relative expression data from 1917 genes as training examples; and the remaining 10% as the test set of examples. The results with this cross-validation were very encouraging (> 97% accuracy on each test set). These results described in Table 2 lead us to believe that there are distinguishable patterns across the five brain regions in response to cocaine that can be modeled as a decision tree. Examples that are misclassified by a good model can also point us toward genes whose expression patterns lie in the boundary areas of the two classes, that is, those patterns that are not entirely describable as belonging to one class versus the other.

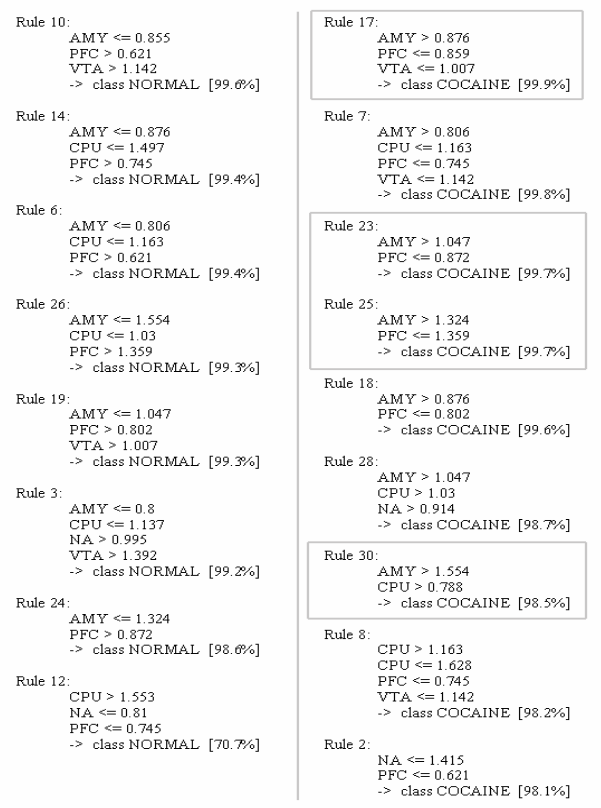
Production Rule Model

The production rules generated by C4.5rules program were carefully studied (Figures 3 and 4). All the rules for a single class appear together and the class subset (the group of rules for a single class) is ordered according to prediction accuracy of the rule; while the rule numbers are based on when they were generated. As in Figure 3, all the rules for the NORMAL class are listed first; then are the rules for the COCAINE class. The rules are applied to each case in the test dataset one by one in the listed order until the case is covered by a rule. The first rule that covers the case will be taken as the operative one since that is the rule with highest classification accuracy. During generation of the decision-tree, all training examples classified by any existing rule are removed from consideration; hence each rule in the rule-set will cover at least one training example not covered by other rules.

Each production rule comprises of a left-hand side and a right-hand side. The left-hand side can contain tests for values of up to five attributes that are normalized gene expression values in five brain tissue samples. The right-hand side contains a classification, which is the name of a target class such as COCAINE or NORMAL. For example, Rule 25 in Figure 3 can be interpreted as follows. The left-hand side of the rule contains tests for values of two attributes which are normalized gene expression values in brain tissue from the Amygdala and Prefrontal Cortex; while the right side is a class name namely COCAINE. A test case that satisfies the left side of this rule is classified as COCAINE. The program also predicts that this classification will be correct for 99.7% of unseen cases that satisfy this rule's left-hand side. This implies that whenever this rule is used to classify an unseen test case that has not been classified by any of the more accurate rules for that class, there is a 99.7% chance that the classification is accurate.

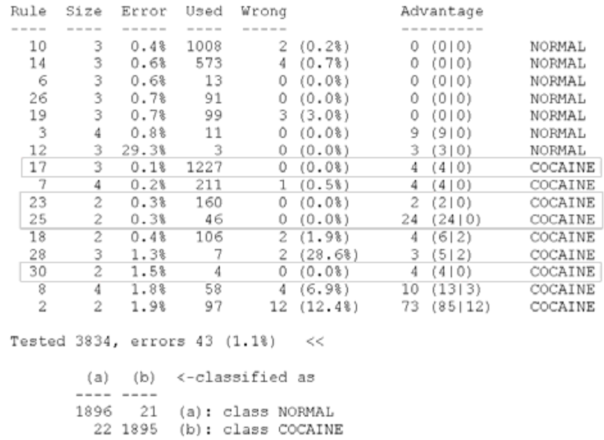
Rule 17, the most generalized rule in the rule set that covered 1227 genes without any misclassification described a pattern of global effects on gene expression in the five different brain regions under cocaine treatment (Figure 4). This pattern showed that genes are up regulated in Amygdala and simultaneously down regulated in the Prefrontal Cortex and to some extent in the Ventral Tegmental Area. We verified this pattern by the result of a clustering study on gene expression of the 1917 genes using GeneSpring 4.1.

Prior to the clustering study, the raw data was firstly normalized sample-wise. Then the relative expression value of each gene in cocaine treated samples was divided by that in normal sample, from each brain region. These normalized relative expression data of a gene indicated the fold change in gene expression between the cocaine treated samples and normal samples from each brain region. This dataset was filtered as described in the METHOD Section. This filtered dataset was therefore referred as the fold change dataset. A gene tree was built from the fold change dataset by using the gene clustering function provided by GeneSpring 4.1. Similarity was measured by spearman correlation; separation ration was set to 0.5; and the



```
Rule 10:
        AMY <= 0.855
        PFC > 0.621
        VTA > 1.142
        ->  class NORMAL  [99.6%]

Rule 14:
        AMY <= 0.876
        CPU <= 1.497
        PFC > 0.745
        ->  class NORMAL  [99.4%]

Rule 6:
        AMY <= 0.806
        CPU <= 1.163
        PFC > 0.621
        ->  class NORMAL  [99.4%]

Rule 26:
        AMY <= 1.554
        CPU <= 1.03
        PFC > 1.359
        ->  class NORMAL  [99.3%]

Rule 19:
        AMY <= 1.047
        PFC > 0.802
        VTA > 1.007
        ->  class NORMAL  [99.3%]

Rule 3:
        AMY <= 0.8
        CPU <= 1.137
        NA > 0.995
        VTA > 1.392
        ->  class NORMAL  [99.2%]

Rule 24:
        AMY <= 1.324
        PFC > 0.872
        ->  class NORMAL  [98.6%]

Rule 12:
        CPU > 1.553
        NA <= 0.81
        PFC <= 0.745
        ->  class NORMAL  [70.7%]

Rule 17:
        AMY > 0.876
        PFC <= 0.859
        VTA <= 1.007
        ->  class COCAINE  [99.9%]

Rule 7:
        AMY > 0.806
        CPU <= 1.163
        PFC <= 0.745
        VTA <= 1.142
        ->  class COCAINE  [99.8%]

Rule 23:
        AMY > 1.047
        PFC <= 0.872
        ->  class COCAINE  [99.7%]

Rule 25:
        AMY > 1.324
        PFC <= 1.359
        ->  class COCAINE  [99.7%]

Rule 18:
        AMY > 0.876
        PFC <= 0.802
        ->  class COCAINE  [99.6%]

Rule 28:
        AMY > 1.047
        CPU > 1.03
        NA > 0.914
        ->  class COCAINE  [98.7%]

Rule 30:
        AMY > 1.554
        CPU > 0.788
        ->  class COCAINE  [98.5%]

Rule 8:
        CPU > 1.163
        CPU <= 1.628
        PFC <= 0.745
        VTA <= 1.142
        ->  class COCAINE  [98.2%]

Rule 2:
        NA <= 1.415
        PFC <= 0.621
        ->  class COCAINE  [98.1%]
```

Figure.3. The set of rules obtained by providing all 3834 training examples to C4.5 and subsequently invoking C4.5rules to create rules from the learned decision tree. These rules provide a plausible, generalized model learned from the training data. The program automatically finds the values for normalized expression levels across one or more brain regions that in combination are predictive of cocaine-treated tissue or normal tissue. Each rule can be used to classify a certain number of training examples to a certain degree of accuracy (indicated in square brackets). The boxed rules did not misclassify any training example. See Figure 1 legend for abbreviations.



```
Rule  Size  Error  Used  Wrong            Advantage
----  ----  -----  ----  -----            ---------
  10     3   0.4%  1008     2  (0.2%)      0  (0|0)      NORMAL
  14     3   0.6%   573     4  (0.7%)      0  (0|0)      NORMAL
   6     3   0.6%    13     0  (0.0%)      0  (0|0)      NORMAL
  26     3   0.7%    91     0  (0.0%)      0  (0|0)      NORMAL
  19     3   0.7%    99     3  (3.0%)      0  (0|0)      NORMAL
   3     4   0.8%    11     0  (0.0%)      9  (9|0)      NORMAL
  12     3  29.3%     3     0  (0.0%)      3  (3|0)      NORMAL
  17     3   0.1%  1227     0  (0.0%)      4  (4|0)      COCAINE
   7     4   0.2%   211     1  (0.5%)      4  (4|0)      COCAINE
  23     2   0.3%   160     0  (0.0%)      2  (2|0)      COCAINE
  25     2   0.3%    46     0  (0.0%)     24  (24|0)     COCAINE
  18     2   0.4%   106     2  (1.9%)      4  (6|2)      COCAINE
  28     3   1.3%     7     2  (28.6%)     3  (5|2)      COCAINE
  30     2   1.5%     4     0  (0.0%)      4  (4|0)      COCAINE
   8     4   1.8%    58     4  (6.9%)     10  (13|3)     COCAINE
   2     2   1.9%    97    12  (12.4%)    73  (85|12)    COCAINE

Tested 3834, errors 43 (1.1%)   <<

       (a)  (b)   <-classified as
      ---- ----
      1896   21   (a): class NORMAL
        22 1895   (b): class COCAINE
```

Figure 4. Coverage of training examples (genes described by their expression values tagged with the type of tissue) by each of the rules in the learned rule set in Figure 3. Consider for example Rule 25, which was used 46 times in classifying the training cases. All of the cases that satisfied the rule's left-hand side did in fact belong to class COCAINE, so this rule did not misclassify any example (Wrong = 0). The advantage of including this rule in the set of learned rules is indicated as 24 (24|0) – this means that if the rule were omitted, 24 cases now classified correctly by this rule would be classified incorrectly, and 0 cases now misclassified by this rule would be correctly classified by the subsequent rules and the default class; the net benefit of retaining the rule is thus 24 = 24 – 0. Other rules could be interpreted similarly. (Quinlan, 1993). The confusion matrix (Kohavi and Provost, 1998) is shown at the bottom of the figure with 1.1% of examples incorrectly classified. The boxed rules are corresponding to the same rules in Figure 3.

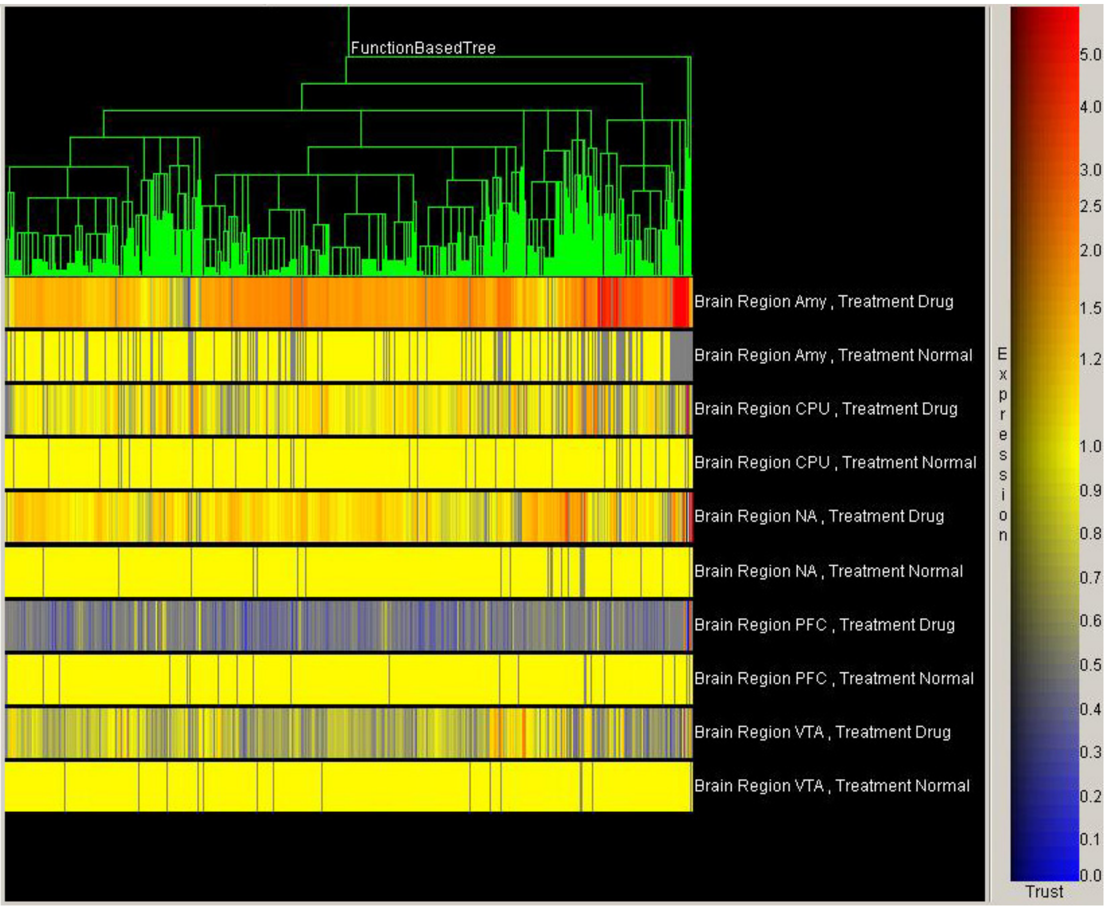


Figure 5. Visualization of the 1917 genes present in all ten samples. Red indicates up-regulated genes and blue indicates down-regulated genes. The expression of these genes are shown in the following order row-wise alternating between cocaine-treated and normal tissue; from Amygdala, Caudate Putamen, Nucleus Accumbens, Prefrontal Cortex and Ventral Tegmental Area. The pattern of global effects of more than half of the genes being over-expressed in Amygdala; and simultaneously under-expressed in the Prefrontal Cortex and to some extent in the Ventral Tegmental Area is clearly visible and is described by Rule 17 in Figure 3.

minimum distance was adjusted to 0.001. The gene tree obtained for visualization (Figure 5) depicted the same pattern described by Rule 17 and provided a strong evidence for the existence of that global effect.

Apart from the model, we listed and annotated genes covered by each rule using Gene Ontology terms from the TIGR Rat Gene Index. The Gene Ontology annotation provides insightful information for a gene categorized as: the cellular component a gene belongs to, its molecular function, and the biological process associated with the gene (The Gene Ontology Consortium, 2000). The Gene Ontology based annotation of the genes covered by each rule and the normalized dataset are made available publicly at http://www.pitt.edu/~chmst40/ratdata/ .

We also studied many of the genes obtained from the rules that do not misclassify any of the training examples (the boxed rules in Figure 3). Many of the genes covered by these rules are implicated in signal transduction mechanisms. For example, the HPC-1 gene covered by Rule 30 was up regulated in both Amygdala and Caudate

Putamen. HPC-1 antigen may play a role in neurotransmitter release from nerve terminals by associating with omega-CgTX-sensitive N-type calcium channel and synaptotagmin (Yoshida et al., 1992). This reinforces the general belief that several signal transduction pathways are activated from the receptor/cell surface to the nucleus and back that regulate the behavioral circuits in the brain leading to the kinds of response seen in animals exposed to drugs.

CONCLUSIONS

Methods for visualization and analysis of large, high throughput gene expression datasets remain an important area of research. In this paper we present such a method based on classical supervised machine learning. In particular, we have successfully used the well known decision tree learning program C4.5 to analyze global effects of acute cocaine exposure in the rat brain. The program was able to generate highly accurate, human-understandable rules that could distinguish cocaine exposed and naïve rat brains on the basis of gene expression data and which were consistent with Spearman correlation based statistical clustering analysis on differently normalized gene expression data from a same set of genes.

Our methodology of normalization and filtering ensured the success of decision tree learning. The importance of normalization and filtering however, is not unique to decision tree learning or this study, it is a critical step first step in analysis of all large gene expression data sets no matter which analytic approach is used.

In order to develop suitable treatment options for drug abuse, science will need to establish whether the effects of drugs such as cocaine are due to local or global changes in gene expression across the various brain regions. Our experiment establishes one such interaction among the Amygdala and the Prefrontal Cortex that could be very useful in this understanding. The Amygdala region of the brain is typically associated with fear and emotion (Agglenton, 2000); while the Prefrontal Cortex is associated with long-term memory, planning and multi-tasking (LeDoux, 1996). The simultaneous down-regulation of more than a thousand genes in the PFC region and up-regulation of the same genes in the Amygdala is a likely contributor to reinstating drug-seeking behavior due to short-term reward or stimulus. Previous studies have implicated both dopaminergic as well as non-dopaminergic systems as being involved in drug abuse and addiction (Lucas et al., 1997; Bhat and Baraban, 1993). This study suggests that there are strong global effects of interaction among brain regions due to the exposure to a drug. Treatment measures designed to counteract these global effects might be more successful than those that consider only local effects of drug exposure.

In conclusion, we have demonstrated that the decision tree learning method can accurately learn from microarray gene expression data and can generate a human-understandable model that can be used for prediction. The model learned from the gene expression data of rat brain with or without cocaine treatments describes a global change of gene expression due to acute cocaine treatment. This model provides evidence for the existence of global effects of cocaine in the rat brain, implicates the interaction of different brain regions under cocaine treatment, and gives insightful information for the treatment of drug abuse.

ACKNOWLEDGEMENTS

We thank Ivet Bahar for her helpful suggestions. We thank John Gilbertson for his careful reading and revising of the manuscript and insightful discussion. We thank Eleanor Feingold for her help on the statistical analysis of the dataset. We thank David Dougall for his help on annotation of genes using TIGR Rat Gene Index. V.G. gratefully acknowledges grant 5T15LM07059-14 from the National Library of Medicine.

REFERENCES

Affymetrix Inc. (2000) Expression analysis technical manual appendix 5.

Aggleton, J.P. (2000) The amygdala. University Press, Oxford.

Bhat, R.V. and Baraban J.M. (1993) Activation of transcription factor genes in striatum by cocaine: role of both serotonin and dopamine systems. J. Pharmacol. Exp. Ther., 267, 496-504.

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, Jr. M., and Haussier, D. (2000) Knowledge-based analysis of microarray gene expression data by using supporting vector machines. Proc. Natl. Acad. Sci. USA, 97, 262-267.

Conway A.R. (199_) GeneSpring Version: Revision 4.1 Redwood City Silicon Genetics.

Eisen, M.B, Spellman, P.T, Brown, P.O., and Botsein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95, 14863-14868.

Hope, B. (1998) Cocaine and the AP-1 transcription factor complex. Ann. Ny. Acad. Sci., 844, 1-6.

Kohavi, R. and Provost, F. (1998) Glossary of terms. editorial for the special issue on applications of machine learning and the knowledge discovery process. 30, 271-274.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nature Genet. 25, 25-29.

LeDoux, J. (1996) The emotional brain. Simon & Schuster. New York.

Lucas, J.J., Segu, L., and Hen, R. (1997) 5-Hydroxytryptamine1B receptors modulate the effect of cocaine on c-fos expression: converging evidence using 5-hydroxytryptamine1B / 1D antagonist GR127935. Mol. Pharmacol., 51, 755-63.

Paxinos, G (1985) Rat Nervous System, Volume 1, Forebrain and midbrain, page 55, Academic Press Inc.

Quinlan, J.R. (1986) Discovering rules by induction from large collections of examples. Machine Learning Journal, 1, 81-106.

Quinlan, J.R. (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers, Inc. San Francisco.

Schadt, E.E., Li, C., Su, C., and Wong, W.H. (2000) Analyzing high-density oligonucleotide gene expression array data. J. Cell. Biochem. 80, 192-202.

Sigenetics Inc. (2001) GeneSpring user manual release 4.1, Appendix G.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA, 96, 2907-2912.

Torres, G., and Horowitz J. M., (1999) Drugs of abuse and brain gene expression. Psychosom. Med., 61, 630-650.

Yoshida A, Oho, C., Omori, A., Kuwahara, R., Ito, T., and Takahashi, M. (1992) HPC-1 is associated with synaptotagmin and omega-conotoxin receptor. J Biol Chem., 267, 24925-8.

BIBLIOGRAPHY

1. Crick FHC: **On protein synthesis**. *Symposia of the Society for Experimental Biology* 1958, **12**(The biological replication of macromolecules):138-163.

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

3. Schena M: **Microarray Analysis**: John Wiley & Songs, Inc.Hoboken, NJ; 2003.

4. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray**. *Science* 1995, **270**(5235):467-470.

5. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays**. *Nature Genetics* 1999, **21**(supplement):10-14.

6. Lipshutz R, Fodor S, Gingeras T, Lockhart D: **High density synthetic oligonucleotide arrays.** *nature Genetics* 1999, **21**(supplement 1):20-24.

7. Meier-Ewert S, Maler E, Ahmadl A, Curtis J, Lehrach H: **An automated approach to generating expressed sequence catalogues**. *nature* 1993, **361**(6410):375-376.

8. Nizetic D, Zehetner G, Monaco AP, Gellen L, Young BD, Lehrach H: **Construction, Arraying, and High-Density Screening of Large Insert Libraries of Human Chromosomes X and 21: Their Potential Use as Reference Libraries**. *PNAS* 1991, **88**(8):3233-3237.

9. Lennon GG, Lehrach H: **Hybridization analyses of arrayed cDNA libraries**. *Trends in Genetics* 1991, **7**(10):314.

10. Milosavljevic A, Zeremski M, Strezoska Z, Grujic D, Dyanov H, Batus S, Salbego D, Paunesku T, Soares MB, Crkvenjakov R: **Discovering distinct genes represented in 29,570 clones from infant brain cDNA libraries by applying sequencing by hybridization methodology**. *Genome Res* 1996, **6**(2):132-141.

11. Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, Jordan BR: **Differential Gene Expression in the Murine Thymus Assayed by Quantitative Hybridization of Arrayed cDNA Clones**. *Genomics* 1995, **29**(1):207.

12. Pietu G, Alibert O, Guichard V, Lamy B, Bois F, Leroy E, Mariage-Sampson R, Houlgatte R, Soularue P, Auffray C: **Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array**. *Genome Res* 1996, **6**(6):492-503.

13. Takahashi N, Hashida H, Zhao N, Misumi Y, Sakaki Y: **High-density cDNA filter analysis of the expression profiles of the genes preferentially expressed in human brain**. *Gene* 1995, **164**(2):219.

14. Zhao N, Hashida H, Takahashi N, Misumi Y, Sakaki Y: **High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression**. *Gene* 1995, **156**(2):207.

15. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D: **Light-Directed, Spatially Addressable Parallel Chemical Synthesis**. *Science* 1991, **251**(4995):767-773.

16. Guo Z, Guilfoyle RA, Thiel AJ, Wang R, Smith LM: **Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports**. *Nucl Acids Res* 1994, **22**(24):5456-5465.

17. Khrapko KR, P. Lysov Y, Khorlyn AA, Shick VV, Florentiev VL, Mirzabekov AD: **An oligonucleotide hybridization approach to DNA sequencing**. *FEBS Letters* 1989, **256**(1-2):118.

18. Lamture JB, Beattie KL, Burke BE, Eggers MD, Ehrlich DJ, Fowler R, Hollis MA, Kosicki BB, Reich RK, Smith SR: **Direct detection of nucleic acid hybridization on the surface of a charge coupled device**. *Nucl Acids Res* 1994, **22**(11):2121-2125.

19. Maskos U, Southern EM: **Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ**. *Nucl Acids Res* 1992, **20**(7):1679-1684.

20. DeRisi JL, Iyer VR, Brown PO: **Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale**. *Science* 1997, **278**(5338):680-686.

21. Harrington CA, Rosenow C, Retief J: **Monitoring gene expression using DNA microarrays**. *Current Opinion in Microbiology* 2000, **3**(3):285.

22. Barczak A, Rodriguez MW, Hanspers K, Koth LL, Tai YC, Bolstad BM, Speed TP, Erle DJ: **Spotted Long Oligonucleotide Arrays for Human Gene Expression Analysis**. *Genome Res* 2003, **13**(7):1775-1785.

23. Dorris D, Nguyen A, Gieser L, Lockner R, Lublinsky A, Patterson M, Touma E, Sendera T, Elghanian R, Mazumder A: **Oligodeoxyribonucleotide probe accessibility on a three-dimensional DNA microarray surface and the effect of hybridization time on the accuracy of expression ratios**. *BMC Biotechnology* 2003, **3**(1):6.

24. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays**. *Nucl Acids Res* 2000, **28**(22):4552-4557.

25. Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, Prokhorova A, Gieser L, Touma E, Lockner R, Tata M *et al*: **An assessment of Motorola CodeLinkTM microarray performance for gene expression profiling applications**. *Nucl Acids Res* 2002, **30**(7):e30-.

26. Wang H-Y, Malek R, Kwitek A, Greene A, Luu T, Behbahani B, Frank B, Quackenbush J, Lee N: **Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays**. *Genome Biology* 2003, **4**(1):R5.

27. Blanchard AP, Kaiser RJ, Hood LE: **High-density oligonucleotide arrays**. *Biosensors and Bioelectronics* 1996, **11**(6-7):687.

28. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR *et al*: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer**. *Nature Biotechnology* 2001, **19**(4):342-347.

29. Wodicka L, Dong H, Mittmann M, Ho M-H, Lockhart DJ: **Genome-wide expression monitoring in Saccharomyces cerevisiae**. *Nat Biotech* 1997, **15**(13):1359.

30. The C.elegans Sequencing Consortium: **Genome Sequence of the Nematode C.elegans: A Platform for Investigating Biology**. *Science* 1998, **282**(5396):2012-2018.

31. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*: **The Genome Sequence of Drosophila melanogaster**. *Science* 2000, **287**(5461):2185-2195.

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. 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**. *PNAS* 2001, **98**(24):13790-13795.

34. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI *et al*: **Diversity of gene expression in adenocarcinoma of the lung**. *PNAS* 2001, **98**(24):13784-13789.

35. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG *et al*: **Gene-expression profiles predict survival of patients with lung adenocarcinoma**. *Nature Medicine* 2002, **8**(8):816-824.

36. Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, Lu C, Woodgett J, Seiden I, Johnston M *et al*: **Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-free Survival**. *Cancer Res* 2002, **62**(11):3005-3008.

37. Chandran U, Dhir R, Ma C, Michalopoulos G, Becich M, Gilbertson J: **Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors**. *BMC Cancer* 2005, **5:45**:doi:10.1186/1471-2407-1185-1145.

38. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP *et al*: **Gene expression correlates of clinical prostate cancer behavior**. *Cancer Cell* 2002, **1**(2):203-209.

39. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, Frierson HF, Jr., Hampton GM: **Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer**. *Cancer Res* 2001, **61**(16):5974-5978.

40. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C, Thomas R, Dhir R, Finkelstein S *et al*: **Gene Expression Alterations in Prostate Cancer Predicting Tumor Aggression and Preceding Development of Malignancy**. *J Clin Oncol* 2004, **22**(14):2790-2799.

41. Kononen J, Bubendorf L, Kallionimeni A, Barlund M, Schraml P, Leighton S, Torhorst J, J Mihatsch M, Sauter G, Kallionimeni O-P: **Tissue microarrays for high-throughput molecular profiling of tumor specimens**. *nature Medicine* 1998, **4**(7):844-847.

42. Simon R, Mirlacher M, Sauter G: **Tissue microarrays**. *Biotechniques* 2004, **36**(1):98-105.

43. Schoenberg Fejzo M, Slamon DJ: **Frozen Tumor Tissue Microarray Technology for Analysis of Tumor RNA, DNA, and Proteins**. *Am J Pathol* 2001, **159**(5):1645-1650.

44. Simon R, Sauter G: **Tissue microarrays for miniaturized high-throughput molecular profiling of tumors**. *Experimental Hematology* 2002, **30**(12):1365.

45. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: **Delineation of prognostic biomarkers in prostate cancer**. *Nature* 2001, **412**(6849):822-826.

46. Barone AD, Beecher JE, Bury PA, Chen C, Doede T, Fidanza JA, McGall GH: **Photolithographic synthesis of high-density oligonucleotide probe arrays**. *Nucleosides, Nucleotides & Nucleic Acids* 2001, **20**(4-7):525-531.

47. McGall GH, Fidanza JA: **Photolithographic synthesis of high-density oligonucleotide arrays**. *Methods in Molecular Biology* 2001, **170**:71-101.

48. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W: **Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists**. *PNAS* 1996, **93**(24):13555-13560.

49. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The Sequence of the Human Genome**. *Science* 2001, **291**(5507):1304-1351.

50. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tomé P, Aggarwal A, Bajorek E *et al*: **A Gene Map of the Human Genome**. *Science* 1996, **274**(5287):540-546.

51. Quackenbush J, Liang F, Holt I, Pertea G, Upton J: **The TIGR Gene Indices: reconstruction and representation of expressed gene sequences**. *Nucl Acids Res* 2000, **28**(1):141-145.

52. Caruthers MH: **Gene Synthesis Machines: DNA Chemistry and its Uses**. *Science* 1985, **230**(4723):281-285.

53. Beaucage SL: **Current protocols in nucleic acid chemistry**: [Hoboken, NJ]: Wiley; 2005.

54. Nishimura M, Yokoi N, Miki T, Horikawa Y, Yoshioka H, Takeda J, Ohara O, Seino S: **Construction of a multi-functional cDNA library specific for mouse pancreatic islets and its application to microarray**. *DNA Research* 2004, **11**(5):315-323.

55. McGall GH, Christians FC: **High-density genechip oligonucleotide probe arrays**. *Advances in Biochemical Engineering-Biotechnology* 2002, **77**:21-42.

56. Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen M-M, Lu G, Fang J, Liu W-M, Ryder T *et al*: **Probe selection for high-density oligonucleotide arrays**. *PNAS* 2003, **100**(20):11237-11242.

57. Zhang W, Shmulevich I, Astola J: **Microarray quality control**. Hoboken, NJ: Wiley-Liss; 2004.

58. Cole K, Truong V, Barone D, McGall G: **Direct labeling of RNA with multiple biotins allows sensitive expression profiling of acute leukemia class predictor genes**. *Nucl Acids Res* 2004, **32**(11):e86-.

59. Gupta V, Cherkassky A, Chatis P, Joseph R, Johnson AL, Broadbent J, Erickson T, DiMeo J: **Directly labeled mRNA produces highly precise and unbiased differential gene expression data**. *Nucl Acids Res* 2003, **31**(4):e13-.

60. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P: **Analysis of Gene Expression in Single Live Neurons**. *PNAS* 1992, **89**(7):3010-3014.

61. Gelder RNV, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH: **Amplified RNA Synthesized from Limited Quantities of Heterogeneous cDNA**. *PNAS* 1990, **87**(5):1663-1667.

62. Petalidis L, Bhattacharyya S, Morris GA, Collins VP, Freeman TC, Lyons PA: **Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis**. *Nucl Acids Res* 2003, **31**(22):e142-.

63. Phillips J, Eberwine JH: **Antisense RNA Amplification: A Linear Amplification Method for Analyzing the mRNA Population from Single Living Cells**. *Methods* 1996, **10**(3):283.

64. Puskás LG, Zvara Á, Hackler Jr. L, Van Hummelen P: **RNA Amplification Results in Reproducible Microarray Data with Slight Ratio Bias**. *BioTechniques* 2002, **32**(6):1330-1340.

65. Richter A, Schwager C, Hentze S, Ansorge W, Hentze MW, Muckenthaler M: **Comparison of Fluorescent Tag DNA Labeling Methods Used for Expression Analysis by DNA Microarrays**. *BioTechniques* 2002, **33**(3):620-630.

66. **Current Protocols in Molecular Biology**: John Wiley & Sons, Inc; 2003.

67. Cox WG, Singer VL: **Fluorescent DNA hybridization probe preparation using amine modification and reactive dye coupling**. *Biotechniques* 2004, **36**(1):114-122.

68. Kaposi-Novak P, Lee JS, Mikaelyan A, Patel V, Thorgeirsson SS: **Oligonucleotide microarray analysis of aminoallyl-labeled cDNA targets from linear RNA amplification**. *Biotechniques* 2004, **37**(4):580-588.

69. Baugh LR, Hill AA, Brown EL, Hunter CP: **Quantitative analysis of mRNA amplification by in vitro transcription**. *Nucl Acids Res* 2001, **29**(5):e29-.

70. Li Y, Li T, Liu S, Qiu M, Han Z, Jiang Z, Li R, Ying K, Xie Y, Mao Y: **Systematic comparison of the fidelity of aRNA, mRNA and T-RNA on gene expression profiling using cDNA microarray**. *Journal of Biotechnology* 2004, **107**(1):19-28.

71. Saghizadeh M, Brown DJ, Tajbakhsh J, Chen Z, Kenney MC, Farber DB, Nelson SF: **Evaluation of techniques using amplified nucleic acid probes for gene expression profiling**. *Biomolecular Engineering* 2003, **20**(3):97-106.

72. Karsten SL, Van Deerlin VMD, Sabatti C, H. Gill L, Geschwind DH: **An evaluation of tyramide signal amplification and archived fixed and frozen tissue in microarray gene expression analysis**. *Nucl Acids Res* 2002, **30**(2):e4-.

73. Dorris DR, Ramakrishnan R, Trakas D, Dudzik F, Belval R, Zhao C, Nguyen A, Domanus M, Mazumder A: **A Highly Reproducible, Linear, and Automated Sample Preparation Method for DNA Microarrays**. *Genome Res* 2002, **12**(6):976-984.

74. Manduchi E, Scearce LM, Brestelli JE, Grant GR, Kaestner KH, Stoeckert CJJR: **Comparison of different labeling methods for two-channel high-density microarray experiments**. *Physiol Genomics* 2002, **10**(3):169-179.

75. Nilsen TW, Grayzel J, Prensky W: **Dendritic Nucleic Acid Structures**. *Journal of Theoretical Biology* 1997, **187**(2):273.

76. Stears RL, Getts RC, Gullans SR: **A novel, sensitive detection system for high-density microarrays using dendrimer technology**. *Physiol Genomics* 2000, **3**(2):93-99.

77. Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J: **A concise guide to cDNA microarray analysis**. *Biotechniques*, **29**(3):548-550.

78. **Affymetrix Expression analysis technical manual**. In.; 2000.

79. Quackenbush J: **COMPUTATIONAL ANALYSIS OF MICROARRAY DATA**. *Nature Reviews Genetics* 2001, **2**(6):418-427.

80. Quackenbush J: **Microarray data normalization and transformation**. *Nature Genetics* 2002, **32**(Supplement):496-501.

81. Chen Y, Dougherty ER, Bittner M: **Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images**. *Journal of Biomedical Optics* 1997, **2**(4):364-374.

82. **Affymetrix Technical Note: Statistical Algorithm Reference Guide.** Affymetrix Inc. Santa Clara, CA; 2001.

83. **Affymetrix White Paper: Statistical Algorithms Descrition Document.** Affymetrix Inc. Santa Clara, CA; 2002.

84. Li C, Hung Wong W: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. *Genome Biology* 2001, **2**(8):research0032.0031 - research0032.0011.

85. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection**. *PNAS* 2001, **98**(1):31-36.

86. Li C, Wong WH: **DNA-Chip Analyzer (dChip)**. In: *The analysis of gene expression data: methods and software.* Edited by Parmigiani G, Garrett E, Irizarry R, Zeger S: Springer, NY; 2003.

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

88. 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**. *Biostat* 2003, **4**(2):249-264.

89. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biology* 2004, **5**(10):R80.

90. **Affymetrix Manual: GeneChip® Expression Analysis Data Analysis Fundamentals by Affymetrix Inc.** [http://www.affymetrix.com/support/technical/manuals.affx]

91. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

92. Morrison DA, Ellis JT: **The Design and Analysis of Microarray Experiments: Applications in Parasitology**. *DNA and Cell Biology* 2003, **22**(6):357-394.

93. Simon R, Radmacher MD, Dobbin K: **Design of studies using DNA microarrays**. *Genetic Epidemiology* 2002, **23**(1):21-36.

94. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U *et al*: **Gene expression profiling identifies clinically relevant subtypes of prostate cancer**. *PNAS* 2004, **101**(3):811-816.

95. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X *et al*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. *Nature* 2000, **403**(6769):503.

96. simon R, Korn E, McShane LM, Radmacher MD, Wright G, Zhao Y: **Design and analysis of DNA microarray investigations**: New York: Springer, 2003; 2003.

97. Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification**. *Journal of the National Cancer Institute* 2003, **95**(1):14-18.

98. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *PNAS* 1998, **95**(25):14863-14868.

99. Aronow BJ, Richardson BD, Handwerger S: **Microarray analysis of trophoblast differentiation: gene expression reprogramming in key gene function categories**. *Physiol Genomics* 2001, **6**(2):105-116.

100. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation**. *PNAS* 1999, **96**(6):2907-2912.

101. Halkidi M, BATISTAKIS; Y, VAZIRGIANNIS; M: **On Clustering Validation Techniques**. *Journal of Intelligent Information Systems* 2001, **17**(2/3):107-145.

102. Ring BZ, Ross DT: **Microarrays and molecular markers for tumor classification**. *Genome Biology* 2002, **3**(5):comment2005.

103. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP *et al*: **Multiclass cancer diagnosis using tumor gene expression signatures**. *PNAS* 2001, **98**(26):15149-15154.

104. Giordano TJ, Shedden KA, Schwartz DR, Kuick R, Taylor JMG, Lee N, Misek DE, Greenson JK, Kardia SLR, Beer DG *et al*: **Organ-Specific Molecular Classification of Primary Lung, Colon, and Ovarian Adenocarcinomas Using Gene Expression Profiles**. *Am J Pathol* 2001, **159**(4):1231-1238.

105. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF, Jr. *et al*: **Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures**. *Cancer Res* 2001, **61**(20):7388-7393.

106. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.

107. van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ *et al*: **A Gene-Expression Signature as a Predictor of Survival in Breast Cancer**. *N Engl J Med* 2002, **347**(25):1999-2009.

108. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT *et al*: **Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification**. *Cancer Res* 2003, **63**(7):1602-1607.

109. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *PNAS* 2001, **98**(9):5116-5121.

110. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C *et al*: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks**. *Nat Med* 2001, **7**(6):673.

111. Dudoit SF, Jane, Speed TP: **Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data**. *journal of the American Statistical Association* 2002, **97**(457):77-87.

112. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines**. *PNAS* 2000, **97**(1):262-267.

113. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data**. *Bioinformatics* 2000, **16**(10):906-914.

114. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression**. *PNAS* 2002, **99**(10):6567-6572.

115. Zhang H, Yu C-Y, Singer B, Xiong M: **Recursive partitioning for tumor classification with gene expression microarray data**. *PNAS* 2001, **98**(12):6730-6735.

116. Wigle D, Tsao M, Jurisica I: **Making sense of lung-cancer gene-expression profiles**. *Genome Biology* 2004, **5**(2):309.

117. Ma C, Michalopoulos GK, Luo J, Becich MJ, Gilbertson JR: **Effects of analysis methods, tissue sampling and tissue processing on the supervised classification of prostate tissue samples using microarray results**. In.: University of Pittsburgh; 2003.

118. Yauk CL, Berndt ML, Williams A, Douglas GR: **Comprehensive comparison of six microarray technologies**. *Nucl Acids Res* 2004, **32**(15):e124-.

119. Tan PK, Downey TJ, Spitznagel EL, Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC: **Evaluation of gene expression measurements from commercial microarray platforms**. *Nucl Acids Res* 2003, **31**(19):5676-5684.

120. **Standardizing global gene expression analysis between laboratories and across platforms**. *Nat Meth* 2005, **2**(5):351.

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

122. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nature Genetics* 2000, **25**(1):25-29.

123. Hosack D, Dennis G, Sherman B, Lane H, Lempicki R: **Identifying biological themes within lists of genes with EASE**. *Genome Biology* 2003, **4**(10):R70.

124. Li J, Pankratz M, Johnson JA: **Differential Gene Expression Patterns Revealed by Oligonucleotide Versus Long cDNA Arrays**. *Toxicol Sci* 2002, **69**(2):383-390.

125. Kothapalli R, Yoder S, Mane S, Loughran T: **Microarray results: how accurate are they?** *BMC Bioinformatics* 2002, **3**(1):22.

126. Kuo WP, Jenssen T-K, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies**. *Bioinformatics* 2002, **18**(3):405-412.

127. Jarvinen A-K, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi O-P, Monni O: **Are data from different gene expression microarray platforms comparable?** *Genomics* 2004, **83**(6):1164.

128. Shippy R, Sendera T, Lockner R, Palaniappan C, Kaysser-Kranich T, Watts G, Alsobrook J: **Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations**. *BMC Genomics* 2004, **5**(1):61.

129. Park PJ, Cao YA, Lee SY, Kim J-W, Chang MS, Hart R, Choi S: **Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference**. *Journal of Biotechnology* 2004, **112**(3):225-245.

130. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: **Independence and reproducibility across microarray platforms**. *Nat Meth* 2005, **2**(5):337.

131. Jarvinen A-K, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi O-P, Monni O: **Are data from different gene expression microarray platforms comparable?** *Genomics* 2004, **83**(6):1164-1168.

132. Bakay M, Chen Y-W, Borup R, Zhao P, Nagaraju K, Hoffman E: **Sources of variability and effect of experimental approach on expression profiling data interpretation**. *BMC Bioinformatics* 2002, **3**(1):4.

133. Dumur CI, Garrett CT, Archer KJ, Nasim S, Wilkinson DS, Ferreira-Gonzalez A: **Evaluation of a linear amplification method for small samples used on high-density oligonucleotide microarray analysis**. *Analytical Biochemistry* 2004, **331**(2):314-321.

134. Gold D, Coombes K, Medhane D, Ramaswamy A, Ju Z, Strong L, Koo JS, Kapoor M: **A comparative analysis of data generated using two different target preparation methods for hybridization to high-density oligonucleotide microarrays**. *BMC Genomics* 2004, **5**(1):2.

135. Chiorino G, Acquadro F, Mello Grand M, Viscomi S, Segir R, Gasparini M, Dotto P: **Interpretation of expression-profiling results obtained from different platforms and tissue sources: examples using prostate cancer data**. *European Journal of Cancer* 2004, **40**(17):2592.

136. Liotta L, Petricoin E: **Molecular profiling of human cancer**. *Nature Reviews Genetics* 2000, **1**(1):48-56.

137. Stamey TA, CALDWELL MC, FAN Z, ZHANG Z, McNEAL JE, NOLLEY R, CHEN Z, MAHADEVAPPA M, WARRINGTON JA: **Genetic Profiling of Gleason Grade 4/5**

Prostate Cancer: Which is the Best Prostatic Control Tissue? *CLINICAL UROLOGY* 2003, **170**(6):2263-2268.

138. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: **Laser Capture Microdissection**. *Science* 1996, **274**(5289):998-1001.

139. Ernst T, Hergenhahn M, Kenzelmann M, Cohen CD, Bonrouhi M, Weninger A, Klaren R, Grone EF, Wiesel M, Gudemann C *et al*: **Decrease and Gain of Gene Expression Are Equally Discriminatory Markers for Prostate Carcinoma: A Gene Expression Analysis on Total and Microdissected Prostate Tissue**. *Am J Pathol* 2002, **160**(6):2169-2180.

140. King HC, Sinha AA: **Gene Expression Profile Analysis by DNA Microarrays: Promise and Pitfalls**. *JAMA* 2001, **286**(18):2280-2288.

141. Luzzi V, Mahadevappa M, Raja R, Warrington JA, Watson MA: **Accurate and Reproducible Gene Expression Profiles from Laser Capture Microdissection, Transcript Amplification, and High Density Oligonucleotide Microarray Analysis**. *J Mol Diagn* 2003, **5**(1):9-14.

142. King C, Guo N, Frampton GM, Gerry NP, Lenburg ME, Rosenberg CL: **Reliability and Reproducibility of Gene Expression Measurements Using Amplified RNA from Laser-Microdissected Primary Breast Tissue with Oligonucleotide Arrays**. *J Mol Diagn* 2005, **7**(1):57-64.

143. Eisen MB, Brown PO: **DNA arrays for analysis of gene expression**. *Methods in Enzymology* 1999, **303**:179-205.

144. Dash A, Maine IP, Varambally S, Shen R, Chinnaiyan AM, Rubin MA: **Changes in Differential Gene Expression because of Warm Ischemia Time of Radical Prostatectomy Specimens**. *Am J Pathol* 2002, **161**(5):1743-1748.

145. Wilson CL, Pepper SD, Hey Y, Miller CJ: **Amplification protocols introduce systematic but reproducible errors into gene expression studies**. *Biotechniques* 2004, **36**(3):498-506.

146. Smith L, Underhill P, Pritchard C, Tymowska-Lalanne Z, Abdul-Hussein S, Hilton H, Winchester L, Williams D, Freeman T, Webb S *et al*: **Single primer amplification (SPA) of cDNA for microarray expression analysis**. *Nucl Acids Res* 2003, **31**(3):e9-.

147. **The New GeneChip® IVT Labeling Kit:Optimized Protocol for Improved Results**. Affymetrix, Santa Clara, CA; 2004.

148. Wang E, Miller LD, Ohnmacht GA, Liu ET, Marincola FM: **High-fidelity mRNA amplification for gene profiling**. *Nature Biotechnology* 2000, **18**(4):457-459.

149. Al-Mulla F, Al-Tamimi R, Bitar MS: **Comparison of two probe preparation methods using long oligonucleotide microarrays**. *BioTechniques* 2004, **37**:827-833.

150. Schneider J, BuneSZ A, Huber W, Volz J, Kioschis P, Hafner M, Poustka A, Sultmann H: **Systematic analysis of T7 RNA polymerase based in vitro linear RNA amplification for use in microarray experiments**. *BMC Genomics* 2004, **5**(1):29.

151. Zhao H, Hastie T, Whitfield M, Borresen-Dale A-L, Jeffrey S: **Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis**. *BMC Genomics* 2002, **3**(1):31.

152. Feldman ALC, N. G. Wang, E. Qian, M. Marincola, F. M. Alexander, H. R. Libutti, S. K.: **Advantages of mRNA Amplification for Microarray Analysis**. *BioTechniques* 2002, **33**:906-914.

153. Polacek DC, Passerini AG, Shi C, Francesco NM, Manduchi E, Grant GR, Powell S, Bischof H, Winkler H, Stoeckert CJ, Jr. *et al*: **Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA**. *Physiol Genomics* 2003, **13**(2):147-156.

154. Gomes LI, Silva RLA, Stolf BS, Cristo EB, Hirata J, Roberto, Soares FA, Reis LFL, Neves EJ, Carvalho AF: **Comparative analysis of amplified and nonamplified RNA for hybridization in cDNA microarray**. *Analytical Biochemistry* 2003, **321**(2):244-251.

155. Nygaard V, Loland A, Holden M, Langaas M, Rue H, Liu F, Myklebost O, Fodstad O, Hovig E, Smith-Sorensen B: **Effects of mRNA amplification on gene expression ratios in cDNA experiments estimated by analysis of variance**. *BMC Genomics* 2003, **4**(1):11.

156. Spiess A-N, Mueller N, Ivell R: **Amplified RNA degradation in T7-amplification methods results in biased microarray hybridizations**. *BMC Genomics* 2003, **4**(1):44.

157. Chetcuti A, Margan S, Mann S, Russell P, Handelsman D, Rogers J, Dong Q: **Identification of differentially expressed genes in organ-confined prostate cancer by gene expression array**. *Prostate* 2001, **47**(2):132-140.

158. Luo JH, Yu YP, Cieply K, Lin F, Deflavia P, Dhir R, Finkelstein S, Michalopoulos G, Becich M: **Gene expression analysis of prostate cancers**. *Molecular Carcinogenesis* 2002, **33**(1):25-35.

159. Sandberg AA: **Chromosomal abnormalities and related events in prostate cancer**. *Human Pathology* 1992, **23**(4):368.

160. Waghray A, Schober M, Feroze F, Yao F, Virgin J, Chen YQ: **Identification of Differentially Expressed Genes by Serial Analysis of Gene Expression in Human Prostate Cancer**. *Cancer Res* 2001, **61**(10):4283-4286.

161. Bostwick D: **Prospective origins of prostate carcinoma: Prostatic intraepithelial neoplasia and atypical adenomatous hyperplasia**. *Cancer* 1996, **78**(2):330-336.

162. Bostwick DG, Shan A, Qian J, Darson M, Maihle NJ, Jenkins RB, Cheng L: **Independent origin of multiple foci of prostatic intraepithelial neoplasia Comparison with matched foci of prostate carcinoma**. *Cancer* 1998, **83**(9):1995-2002.

163. Bartels PH, Montironi R, Duval da Silva V, Hamilton PW, Thompson D, Vaught L, Bartels HG: **Tissue architecture analysis in prostate cancer and its precursors: An innovative approach to computerized histometry**. *European Urology* 1999, **35**(5-6):484-491.

164. Mairinger T, Mikuz G, Gschwendtner A: **Nuclear chromatin texture analysis of nonmalignant tissue can detect adjacent prostatic adenocarcinoma**. *Prostate* 1999, **41**(1):12-19.

165. Montironi R, Diamanti L, Pomante R, Thompson D, Bartels PH: **Subtle changes in benign tissue adjacent to prostate neoplasia detected with a Bayesian belief network**. *Journal of Pathology* 1997, **182**(4):442-449.

166. Quinlan JR: **Induction of Decision Trees**. *Machine Learning* 1986, **1**(1):81.

167. Quinlan J: **C4.5: programs for machine learning.** San Mateo, CA: Morgan Kaufmann Publisher; 1993.

168. Quinlan JR: **C4.5: programs for machine learning**: Morgan Kaufmann Publisher, Inc.; 1993.

169. Freund Y, Schapire R: **Experiments with a new boosting algorithm.** In: *Machine Learning: Proceedings of Thirteenth International Conference.* San Francisco, CA: Morgan Kauffman Publishers.; 1996: 148-156.

170. Prakash K, Pirozzi G, Elashoff M, Munger W, Waga I, Dhir R, Kakehi Y, Getzenberg RH: **Symptomatic and asymptomatic benign prostatic hyperplasia: Molecular differentiation by using microarrays**. *PNAS* 2002, **99**(11):7598-7603.

171. Kerr MK, Churchill GA: **Statistical design and the analysis of gene expression microarray data**. *Genetical Research* 2001, **77**(2):123-128.

172. **Affymetrix GeneChip® Operating Software**: Affymetrix Inc. Santa Clara, CA; 2005.

173. Johnson K, Lin S: **QA/QC as a Pressing Need for Microarray Analysis: Meeting Report from CAMDA'02**. *BioTechniques* 2003, **Mar.**:Suppl: 62-63.

174. **CodeLink Gene Expression System: Manual Labeled cRNA Target Preparation**. In. GE Healthcare, Piscataway, NJ; 2004.

175. **Affymetrix GeneChip Expression Analysis Technical Manual.** In. Affymetrix, Santa Clara, CA; 2001.

176. **Technical Note: The New GeneChip IVT Labeling Kit: Optimized Protocol for Improved Results**

177. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. *Genome Biology* 2001, **2**(8):research0032.0031-0032.0011.

178. Watson JD, Crick FHC: **Molecular structure of nucleic acide. A structure for deoxyribose nucleic acid.** *nature* 1953, **171**:737-738.

179. Sastry SS, Ross BM: **Nuclease Activity of T7 RNA Polymerase and the Heterogeneity of Transcription Elongation Complexes**. *J Biol Chem* 1997, **272**(13):8644-8652.

180. Yauk CL, Berndt ML, Williams A, Douglas GR: **Comprehensive comparison of six microarray technologies**. *Nucl Acids Res* 2004, **32**(15):e124-.

181. Members of the Toxicogenomics Research Consortium: Weis BK: **Standardizing global gene expression analysis between laboratories and across platforms**. *Nat Meth* 2005, **2**(5):351.

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

183. Chou C-C, Chen C-H, Lee T-T, Peck K: **Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression**. *Nucl Acids Res* 2004, **32**(12):e99-.

184. Relogio A, Schwager C, Richter A, Ansorge W, Valcarcel J: **Optimization of oligonucleotide-based DNA microarrays**. *Nucl Acids Res* 2002, **30**(11):e51-.

185. Sartor M, Schwanekamp J, Halbleib D, Mohamed I, Karyala S, Medvedovic M, Tomlinson CR: **Microarray results improve significantly as hybridization approaches equilibrium**. *Biotechniques* 2004, **36**(5):790-796.