The Application of the Permutation Test in Genome Wide Expression Analysis

by

Timothy Chan
B.Sc., UBC, BC, Canada, 2002

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Master of Science
in
THE FACULTY OF GRADUATE STUDIES
(Computer Science)

The University of British Columbia
March 2006
© Timothy Chan, 2006
Abstract

We are now in a new era. The recent completion of the entire sequence of the human genome and high-throughput gene expression technologies has transformed the era of molecular biology to the era of genomics. Already, such technologies are showing great promise in disease classification and gene targets. However, like any new exciting technology, great promise and anticipation can lead to wasted resources and false hope. It is critical that we recognize the experimental limitations of these new technologies and most importantly, hidden problems must be addressed.

The primary goal of a high-throughput gene expression experiment is to identify genes of interest that are differentially expressed between two sample groups. This thesis addresses two key issues that have hindered high-throughput gene expression technologies. The first is the sample size issue. Small sample sizes affect statistical confidence and are much more sensitive to outliers. Thus, we show that by using a nonparametric statistical test known as the permutation test, we can achieve higher accuracy than conventional parametric statistical tests such as the $t$-test. The second issue we address is the use of housekeeping genes for normalization of mRNA levels. It is well known that many biological experiments require a set of reference genes that are highly expressed and constant from sample to sample.
The choice of reference genes is critical as the wrong choice can have dire effects on subsequent analyses. To address this issue, we developed a methodology based on SAGE, which is a genome wide expression technology that does not require normalization. Our results suggest that reference genes chosen by our methodology are more appropriate for mRNA normalization than the standard set of housekeeping genes. Furthermore, our results suggest that reference genes are more effective if chosen in a tissue-specific manner.
Contents

Abstract ii

Contents iv

List of Tables vii

List of Figures viii

Acknowledgments x

Dedication xi

1 Introduction 1

1.1 The Genomic Era 1

1.2 Challenges in Genome Wide Expression Analysis 2

1.2.1 Sample Size 2

1.2.2 Use of Housekeeping Genes for Normalization 4

1.3 Contributions 5

1.3.1 Differential Gene Expression Finder 5

1.3.2 Constant Gene Expression Finder 6

1.4 Outline 7
# Table of Contents

4 Constant Gene Expression Finder  
4.1 Related Work ............................................. 63  
4.2 Novel Methodology ...................................... 64  
4.3 Validation ................................................. 68  
4.4 Importance of Tissue Specificity ....................... 70

5 Conclusion and Future Directions ....................... 74  
5.1 Conclusion ................................................ 74  
5.2 The Future of Gene Expression Analysis ............... 76

Bibliography .................................................. 79

Appendix A Central Dogma of Biology ..................... 84

Appendix B Genome-wide Expression Technologies ......... 86  
B.1 DNA Microarrays .......................................... 86  
B.2 SAGE Technology .......................................... 89

Appendix C Validation of Differentially Expressed Genes .... 91  
C.1 Validation of Differentially Expressed Genes of Breast Data .......... 91  
C.2 Validation of Differentially Expressed Genes of Breast Data .......... 96

Appendix D Normalization of Expression Using Permutation Test  
and SAGE (NEPS) Reference Genes  
D.1 Breast-SAGE Reference Genes .......................... 100  
D.2 Brain-SAGE Reference Genes ............................ 100  
D.3 Lung-SAGE Reference Genes ............................. 100  
D.4 All-SAGE Reference Genes ............................... 100
List of Tables

3.1 Statistical Comparison of the Top 30 Genes for Breast Data . . . . . . 38
3.2 Statistical Comparison of the Top 30 Genes for Brain Data . . . . . . 38

4.1 Comparison of the Top 40 Most Differentially Expressed Gene of Mi-
icroarray Data Normalized by two Approaches . . . . . . . . . . . . . . . . . 71

D.1 Breast-SAGE Reference Genes . . . . . . . . . . . . . . . . . . . . . . . . 101
D.2 Brain-SAGE Reference Genes . . . . . . . . . . . . . . . . . . . . . . . . 102
D.3 Lung-SAGE Reference Genes . . . . . . . . . . . . . . . . . . . . . . . . 103
D.4 All-SAGE Reference Genes . . . . . . . . . . . . . . . . . . . . . . . . . 104
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Experimental steps in Serial Analysis of Gene Expression (SAGE)</td>
<td>22</td>
</tr>
<tr>
<td>2.2</td>
<td>Illustrates that every additional sample increases the power of the test exponentially</td>
<td>25</td>
</tr>
<tr>
<td>3.1</td>
<td>Illustrates that the higher the number of iterations, the more stable the list becomes</td>
<td>36</td>
</tr>
<tr>
<td>3.2</td>
<td>Diagram of a typical box and whisker plot</td>
<td>44</td>
</tr>
<tr>
<td>3.3</td>
<td>Graphical Representation of Steps of Analysis for CIS Specific &quot;Genes&quot;</td>
<td>45</td>
</tr>
<tr>
<td>3.4</td>
<td>Boxplots showing gene expression patterns of genes with high CIS expression</td>
<td>47</td>
</tr>
<tr>
<td>3.5</td>
<td>CIS genes that are down-regulated</td>
<td>48</td>
</tr>
<tr>
<td>3.6</td>
<td>A Model of the Basic Structure of Desmosomes</td>
<td>50</td>
</tr>
<tr>
<td>3.7</td>
<td>Expression pattern DSG3 in normal(N), metaplasia(M), CIS (C), and invasive (I) stages</td>
<td>52</td>
</tr>
<tr>
<td>3.8</td>
<td>Expression pattern of DSC2 and DSC3 in normal(N), metaplasia(M), CIS(C) and invasive(I) stages</td>
<td>53</td>
</tr>
<tr>
<td>3.9</td>
<td>Expression patterns of armadillo proteins in normal(N), metaplasia(M), CIS(C), and invasive(I)</td>
<td>53</td>
</tr>
</tbody>
</table>
3.10 Gene expression patterns of a desmoplakin (plakin family) in normal(N), metaplasia(M), CIS(C), and invasive(I) .... 55
3.11 Gene Expression Patterns of Candidate Invasive Biomarkers .... 56
3.12 Gene Expression Patterns of Candidate Invasive Biomarkers .... 58
3.13 Boxplots showing genes that are down regulated in lung cancer regardless of stage . 59
3.14 Boxplots showing gene expression pattern for CIS and Invasive specific stages ... 61
4.1 Raw Count versus Variability on Lung SAGE data .......... 67
4.2 Table showing relative comparison of validation results. .... 69
4.3 Boxplots that depict the importance of tissue specificity .... 73
Acknowledgments

There are a number of people I would like to acknowledge. First and foremost, my supervisor who I am most indebted to for his guidance and wisdom. Without him, I would never have finished this thesis. Second, I am forever grateful to Wan Lam, Stephen Lam and Calum MacAulay at the BC Cancer Research Centre for their invaluable contributions including providing me the data set to work with. Third, I would like thank my mom and family for their financial and emotional support.

TIMOTHY CHAN

The University of British Columbia
March 2006
To my father.
Chapter 1

Introduction

1.1 The Genomic Era

The Human Genome Project was launched in 1988 in a world effort to decode the entire human DNA. The main goals of the project included identifying all 20 to 30 thousand human genes and to accurately sequence the entire human DNA. By February 15, 2001, the International Human Genome Mapping Consortium Published its first physical map in *Nature* that covered approximately 96% of the human genome [1]. Shortly after, on April 14, 2003, the project was officially complete two years ahead of schedule. The project was accomplished by *shot gun sequencing* with the aid of a physical map of the genome. This great scientific accomplishment was only the birth of the genomic era. Today researchers are actively using genome research to fight common diseases such as diabetes, cancer, and heart disease. Decade-old technologies such as DNA microarrays and Serial Analysis Of Gene Expression (SAGE) are finally coming of age as they are showing great promise as tools to detect diseases in their early stages and to aid discovery of candidate drug targets. However, due
to complexities of the biology and the experimental procedures of these techniques, we currently cannot use these technologies as standard tests in hospitals and clinics. In this thesis, we address some of these challenges.

1.2 Challenges in Genome Wide Expression Analysis

In the past decade, there has been an exponential growth in biological experimental data. Two primary technologies responsible for this rise are Serial Analysis of Gene Expression (SAGE) and the more popular DNA Microarrays. These two technologies were developed around the same time and have revolutionized genome-wide expression analysis. One of the key goals of these technologies is to allow the identification of genes that are either up or down regulated in specific conditions (i.e. diseases or environmental conditions). A second goal is to map these expression levels to biological pathways so that biologists may begin to formulate hypotheses on the condition of interest.

While collecting data poses some challenges, it is clear that the biggest challenge lies in the analysis of the collected data. In this thesis, we address two main issues:

1.2.1 Sample Size

For every genome wide expression experiment, samples, which include tissues or cell cultures, are processed to extract the expressed mRNA molecules. These molecules are then collected and counted via the microarray or SAGE method. Often, samples are limited and thus, are subject to poor statistical significance. This is due to two primary reasons:
1. COST: Gene expression technologies are quite expensive and thus, the number of samples are often hindered by cost. For instance, typical Affymetrix DNA microarray chips cost around 550 USD each. Moreover, there is still the cost of obtaining the sample, the cost of microarray analysis software, and the cost of the scanner. A SAGE experiment is even more costly as sequencing is an expensive process. In fact, to sequence the entire genome today, it would theoretically cost around 20 million dollars. Furthermore, a single SAGE library can cost upwards of $50,000 depending on how deep the sample is sequenced.

2. RARE SAMPLES: Sometimes, even if funding is not an issue, the samples are simply rare. For example, with our collaboration with BC Cancer Research Centre (BCCRC), we studied the early detection of lung cancer. Lung cancer is notorious for being difficult to detect in its early stages. One main reason for this is that lung cancer is usually symptomless until its late invasive (INV) stages. Thus, early lesions known as Carcinoma In Situ (CIS) are rare encounters. Even after a few years of data collection, the number of CIS samples in BCCRC’s cohort has only reached about 10. Failure to detect lung cancer at its early stages leads to poor prognoses. Thus, survival rates of lung cancer patients is among the lowest of cancer types.

Small samples sizes are a major problem because of the high rate of false positives in genome wide expression experiments. In fact, both the SAGE and the microarray method require some sort of additional biological validation (ie. RT-PCR) before they are accepted as evidence. Thus, it is crucial that we minimize the amount of false positives as each of these validation procedures take time and
money.

1.2.2 Use of Housekeeping Genes for Normalization

Normalization is a preprocessing step that is required in many biological experiments to balance the non-biological differences between different samples so that they can be compared. Usually, housekeeping genes are used as references because they are believed to be expressed at a constant and abundant level (hereafter referred as the constancy requirement) regardless of whether the cell is active, dividing, or simply idle. Experiments such as QRT-PCR and DNA microarrays rely on such genes for normalization. Microarrays in particular are extremely sensitive to normalization because of the many sources of error and the nature of the data. The data is intensity-based and represents the number of transcripts that naturally hybridize to a DNA microarray chip. Interpreting intensity data as mRNA count numbers a controversial issue. Furthermore, DNA microarrays are also sensitive to many other factors such as the hybridization rate in varying environments, the spot intensity issues, and many others described in the next chapter. These variations make inter-chip comparisons difficult. Thus, microarray data must be observed with much scrutiny as the intensity value of say 2 on chip A may not be the same as the intensity value of 2 on chip B. Normalization is a critical step so that non-biological and biological differences can be distinguished.

In the past few years, it has become increasingly apparent that the commonly used housekeeping genes, such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and beta actin (ACTB), violate the constancy requirement [35] [36] [37] [38]. The problem is that normalizing to such genes can skew data and affect any subsequent analyses.
1.3 Contributions

To address the above problems, our method uses two major components. First, we use a nonparametric statistical test known as the permutation test to deal with the small sample size problem. The permutation framework has the following benefits:

1. It is designed for small sample sizes.
2. It does not make any assumption about the underlying population (a non-parametric test).
3. It is robust to outliers.

Second, we use SAGE (Serial Analysis of Gene Expression) for the following benefits:

1. It is an open method so there is absolutely no bias. That is, we do not need to know what genes we are looking for beforehand.
2. It does not possess the same normalization issues as DNA Microarrays as SAGE is based on sequencing fragments of transcripts which are counted in absolute terms.

As will be shown in Section 1.3.1 and Section 1.3.2, the contributions of this thesis are based on different ways of applying the permutation framework to SAGE in order to deal with the problems stated in Section 1.2.

1.3.1 Differential Gene Expression Finder

A common experiment in genome wide expression analysis is to find differentially expressed genes between diverse samples. For example, we may want to find the
genes that are turned on in tumor tissues but turned off in normal tissues. A popular way to do this is to compare the mean cancer gene expression levels to the mean normal gene expression levels. In this thesis, we will demonstrate the ability of the permutation framework to successfully identify differentially expressed genes between two diverse samples. To do this, we conducted two experiments. The first experiment shows that the permutation framework appears to be much more effective in identifying target candidate genes than the typical two sample t-test of unequal variance. The data used here are the breast and brain SAGE data from NCBI’s SAGEmap site (http://www.ncbi.nlm.nih.gov/SAGE/). The second experiment further shows that the permutation framework is effective in identifying biologically significant genes which warrant further biological investigation. That is, we demonstrate the hypothesis generating power of the permutation test applied to SAGE data. Here, we use a cohort of Lung cancer SAGE libraries at various stages (ie. Normal, Metaplasia, CIS, and Invasive) provided by our research partners at BC Cancer Agency. More specifically, using these set of libraries, we identified candidate biomarkers of lung cancer.

1.3.2 Constant Gene Expression Finder

Interestingly, the permutation framework is also useful for doing the reverse. As will be seen in Chapter 4, it is also useful in identifying constantly expressed genes. Experimentally, this has never been done on a global scale with the exception with microarrays. However, extracting constantly expressed genes from a method that requires normalization itself is arguably inappropriate. Thus, this thesis proposes to use the permutation framework on SAGE to find constantly expressed genes and then to use these genes as references (ie. housekeeping genes). Our results suggest
that our set of SAGE selected reference genes are much more appropriate than the commonly used housekeeping genes. As will be shown in Chapter 4, it may also be important that these reference genes must be tissue specific. That is, the reference genes for the lung may not be the same reference genes as for breast tissue.

1.4 Outline

The outline of this thesis is as follows. In the next section, we first summarize some basic biological concepts. We then go into some details of the two most popular gene expression technologies; Serial Analysis of Gene Expression (SAGE) and DNA Microarrays. Lastly, we describe the permutation test in detail including the steps/algorithm. In Chapter 3, we describe two experiments that use the permutation test as a differential gene expression finder. The first experiment analyzes brain and breast SAGE data and compares the permutation test’s performance with the standard two-sample of unequal variance \( t \)-test. The second experiment analyzes Lung SAGE data provided by BC Cancer Agency and compares the different stages of lung carcinoma in attempt to discover potential biomarkers. In Chapter 4, we describe a novel normalization procedure that identifies constant expressed genes with the permutation test. Finally, we conclude with a brief summary and suggest further directions.
Chapter 2

Background

2.1 The Central Dogma of Biology

It is well known that our bodies are composed of units called genes that are responsible for the various physical and mental characteristics that define us. Genes are found in our DNA and contain templates for specialized messenger molecules known as RNA. RNA in turn contains information used to make the biological worker molecules known as proteins. Proteins have a variety of roles including signaling, immunity, physical structure, and inhibition/promotion of growth. The central dogma of biology refers to the transition between DNA to RNA (process known as transcription) to protein (a process known as translation).

While the human cell is estimated to contain about 30,000 genes, the average cell only has about 20 percent of their genes turned on. From a molecular perspective, the difference between muscle cells, skin cells, nerve cells are simply the different genes that are turned on or off in the particular cell type. This differential expression profile defines the type of cell. A mature cell’s genetic make-up can also change
when there are changes in the environment to adapt and survive. For instance, heat shock proteins are proteins that are elevated in stressed conditions such as an increase in temperature or other environmental changes. One of the most important roles of these proteins is to assist protein-protein conformations and maintain those conformations. In environments where protein degradation is likely, the RNA of heat shock proteins increase. At the same time, the RNA of metabolic proteins is likely to decrease as temperatures are not optimal for such activity. As it turns out, diseased cells also have different genetic signatures from healthy cells. Thus, one of the key purposes of gene expression analysis is to identify key genes that are responsible for the onset and phenotype of the disease. Such knowledge is extremely helpful when designing treatments.

2.2 Genome Wide Expression Technologies

Although we truly would like to analyze the protein levels in cells, it is currently extremely difficult due to various reasons. First, all proteins have their own behavior and many are sensitive to degradation. Thus, even if we can extract the protein, often they would degrade so fast that we would not be able to measure their expression accurately. Second, proteins are often interacting with other genes often involving chemical and physical interactions between other proteins forming large multi-protein complexes. This field of study is known as proteomics and is much more complicated than genomics. Protein microarrays are in the works but are years away from being as wide-scale, accurate and reproducible as gene expression technologies. Thus, for the past decade, researchers have been mostly analyzing gene expression indirectly. Fortunately, these technologies have proved to be quite useful in extracting interesting information and most importantly in developing hypotheses
for further validation and study.

Gene expression analysis began about a decade ago when two genome wide expression technologies were introduced that revolutionized functional genomics by allowing cross-tissue comparisons of expression profiles; serial analysis of gene expression (SAGE) and DNA microarrays. Both methods have been widely used to elucidate complete gene expression profiles. Over the years, these methods have been enhanced via statistical/mathematical and experimental techniques to account for their various weaknesses. For instance, long-SAGE was developed to allow cloning of 20 nucleotide SAGE tags leading to improvements in tag to gene mappings [9], while MicroSAGE was developed to allow SAGE to be done when tissue material is scarce [10]. In DNA microarrays, various techniques have been developed to account for its sensitivity to normalization such as the use of control genes [30] [29]. These techniques will be discussed later in the next section.

2.2.1 DNA Microarrays

DNA microarrays were developed to facilitate the understanding of how a cell coordinates expression of thousands of genes in different conditions (ie. diseased state). When a microarray experiment is performed, a “snap shot” of a cell’s global gene expression levels is taken giving researchers a large list of genes that are currently activated or suppressed. This technology has allowed scientists to conduct many experiments inconceivable just over a decade ago. For example, researchers can now compare the genetic profile of diseased and normal cells and allow extraction of possible drug targets and key genes that may have caused the onset of the disease. Thus, contributing to both early detection and prevention.

In the following section, we describe two of the more popular types of DNA
microarrays and describe their similarities/differences and strengths/weaknesses. We then describe the experimental steps required in a DNA microarray experiment.

**Spotted Microarrays and Affymetrix Oligonucleotide Microarrays**

Spotted microarrays require molecules known as cDNA. cDNAs are reverse transcribed mRNA molecules and thus are free of premature mRNA sequences (ie. exons). Spotted microarrays are made up of several thousand cDNA probes that are fixed on a slide. In general, spotted microarrays are more popular than commercially produced arrays (ie. Affymetrix slides) for two main reasons. First, these types of chips are relatively inexpensive compared to commercially produced chips. Second, spotted microarrays are very flexible since they allow scientists to design their own arrays and spot any types of DNA of interest on to their customized slide.

Unfortunately, spotted microarrays have a few pitfalls. One major drawback is the high error rate during the spotting process. Many of these errors are due to variations in experimental procedures and environments that the experiment is conducted. In fact, it is well known that results from one lab will often vary from another lab and even vary from experiment to experiment. Also, most spotted arrays use cDNA as probes and consequently, require known DNA for the PCR process to produce the probes. In spite of the whole genome being sequenced, every human transcript is still unknown. Lastly, this method results in long DNA fragments on the spotted slide and thus, increases the chances of probes hybridizing to similar cDNA (gene) fragments. Thus, it is virtually impossible to detect small differences between DNA fragments (SNPs cannot be detected).

The second common type of microarray is primarily produced by the company, Affymetrix. These DNA chips differ from spotted microarrays in two distinct
manners. First, this technology is based on hybridization of mRNA samples to small, high-density arrays containing tens of thousands of synthetic oligonucleotides (synthetic DNA). This allows for detection of SNPs (single nucleotide polymorphism) and other small features of DNA (ie. specialized regions such as enhancers, promoters, alternative spliced genes etc). The second major difference is that only one sample is hybridized to the chip at a time while spotted microarrays require a control and experimental sample to be competitively hybridized onto the chip. Thus, Affymetrix chips produce data of absolute intensities while spotted microarrays produce data of relative intensities.

Like spotted microarrays, these chips do have their share of disadvantages. First, Affymetrix chips and the equipment to read them are costly. Currently, on the Affymetrix website (www.affymetrix.com), a human gene chip goes for around 550 USD and the basic Fluidics station and scanner is over $100,000. As well, a custom chip would be even more costly. Second, often the signals generated from different probes for the same gene vary in magnitude (sometimes as much as two-fold). Thus, combining these signal intensities into one intensity to represent and estimate of the abundance of the gene is an issue. Furthermore, since Affymetrix currently only produces DNA chips with only one DNA sample, it has become a major challenge to normalize between the experimental and the control. These challenges will be discussed in the following section.

Experimental Steps of Affymetrix DNA Microarrays

A simple microarray experiment requires 6 basic steps:

1. Sample preparation: The process of obtaining the cleanest sample possible.
   
   This process is the most sensitive to errors in microarray experiments as it is
very difficult to obtain clean samples that are subject to various environmental factors.

2. Array fabrication: The primary concern in this process is to develop arrays with high precision with both lower cost and time to manufacture.

3. Hybridization of the sample to the array: In this step, researchers work on developing ways to increase the efficacy of hybridization.

4. Scanning and image analysis: Much of this part of the analysis deal with issues such as spot location, background correction and intensity assignments.

5. Normalization/Data preprocessing: The goal of normalization is to distinguish non-biological differences from biological differences.

6. Data Analysis: The objective here is to find interesting gene expression patterns and groups using various statistical and data mining methods.

Out of all these steps, normalization is arguably the most useful strategy to correct for non-biological systemic errors so that comparisons between arrays can be performed appropriately. However, in spite of years of research, the normalization of microarrays remain non-standardized. Below we describe this process in detail while further details of the remaining steps are discussed in detail in Appendix B.

**Normalization of DNA microarray chips**

Normalization in gene expression experiments is a broad term for a process of transforming mRNA expression values so that the non-biological differences between experiments is balanced so that the real biological differences between experiments can be observed. This procedure is required in experiments such as RT-PCR and
DNA microarrays. It should be stressed that any form of data manipulation does in itself introduce unwanted noise to the expression measures. Normalization is not a miracle step. The technique can only correct for minor variables but if the data is poor, no amount of normalization can save it.

Although many advances in normalization of DNA microarrays have been made in the past decade, there is still no general consensus on which approach is the most appropriate. Unlike the standardized normalization process in RT-PCR, DNA microarray normalization is much more complicated. One primary reason is the numerous non-biological variables involved in a DNA microarray experiment. For example, a common variable that must be dealt with is the variation in the number of cells in different samples since this would mean that we may have unequal quantities of starting DNA from array to array. Another common problem in microarrays are spatial effects. These problems range from the unequal distribution of solvent across the surface of the array to the quality of washing off non-hybridized mRNA. Furthermore, hybridization efficiency is another common issue. The strength of a hybridized DNA can be affected by the GC versus AT content. GC pairings tend to be stronger as they contain 3 hydrogen bonds while AT pairings are weaker with 2 hydrogen bonds. Also, since different genes often have homologous regions, cross hybridization is a possibility. Moreover, statisticians also must take experimental bias into account. For example, one experimenter may decide to add fluorescent labels for all their chips in one step or separately for each chip. The above is only a subset of issues that need to be considered when analyzing DNA microarrays. In the next section, we will describe some ways to deal with some of these common variables.

Before describing some of these normalization options, it is important to
note that the normalization process is applied after image analysis and before other subsequent data mining analyses including supervised and unsupervised clustering analyses (e.g., hierarchical clustering, K-means clustering, or self-organizing maps).

Some Normalization Options

Over the past decade, many normalization methods have been proposed, many of which were designed for specific conditions of an experiment being run. That is, depending on what the experiment entails, there are choices to be made to yield the most optimal results. For brevity, we will only describe a few of the more popular and newer methods. The first step in normalization is to come up with an appropriate set of genes. Yang et al. suggested three basic approaches: (1) Use all genes on the array. (2) Use constantly expressed genes as references (i.e., housekeeping genes). (3) Use control genes [29] [30].

All Genes on Array

This type of normalization is a simple approach and is based on the following assumption. Since only a small fraction of genes will be up or down regulated in the samples, on average, the number of up and down regulated genes ought to counter each other's effect on the total RNA count. In other words, there should be equal weights of RNA for all samples and thus, the number of RNA molecules should also be roughly the same from experiment to experiment. This assumption is valid for genome wide expression experiments covering the entire genome because we are dealing with thousands of genes where only a handful of them are differentially expressed. The problem with this method is that most microarray chips do not cover the entire genome (especially for homo sapiens). In addition, this method also makes
the assumption that the total mRNA remains constant under different experimental conditions. Consequently, this method would only really be useful if the chips that one compares have similar expression patterns and invalid with chips with vastly different genetic signatures [29]. For instance, the genetic profile of tumor cells are known to be vastly different than corresponding normal cells since tumor cells are essentially in hyperdrive.

A variant of using all genes is normalizing to the total or ribosomal RNA. This method is based on the assumption that ribosomal RNA is constitutively expressed in all cells. Also, since ribosomal RNA consists of 90% of the total RNA in a typical cell, it was thought that it could be used as a normalization standard for small fluctuations in mRNA levels. However, studies have shown that this underlying assumption is not correct [36]. The total RNA produced in a cell can vary a lot from cell to cell and it is dependant on many factors including the conditions the cell is in and the cell type.

Using Constantly Expressed Genes for Normalization

Normalizing to a set of constantly expressed genes is also a popular method and used in Affymetrix's 133 series chips. Instead of using the whole set of genes for normalization, a small subset of genes are used, where chips are re-scaled so that the average values of each housekeeping gene are equal across chips. These genes are chosen based on the biological belief that they are required for a cell to function and live. They are typically picked by selecting for abundant expression in all developmental stages of the cell and are assumed to be non-regulated over various conditions. However, as explained in detail later, these biological assumptions may be too strong. In fact, in recent years, there have been many studies evaluating the
variability of these housekeeping genes [35] , [36] , [40] , [38] , [39] . Not surprisingly, we observed similar trends in our own results.

In the past few years, many have proposed ways to improve on this method. For instance, Kepler et al. proposed the choice of housekeeping genes to be made by looking for a stable background pattern of activity on the microarray. Using this assumption, they can derive a transcriptional core from this stable background pattern (identified statistically for each experiment). This method is referred to as normalization by self-consistency and uses the following steps to acquire the core.

First, all genes are designated as core genes. Next, the whole chip is normalized to all the genes on the chip. After this step, all the genes that have not changed from the normalization process are kept. This process is repeated until the previous transcriptional core and the next transcriptional core are equal [31] . Another method of finding constantly expressed genes proposed by Tseng et al. is using a rank invariance gene approach. This approach attempts to find genes that are non-differentially expressed by first separately computing the intensities of fluorescent labels (of different colors). These labels are then compared and if the ranks of the two differ by some arbitrarily set threshold and the rank of the averaged intensity is not among the highest $1$ ranks or lowest $1$ ranks, the gene is classified as a non-differentially expressed gene [32] .

The problem with these two methods is that they are based on using on-chip data to find non-differentially expressed genes. This, in itself, introduces a circular argument since these so called constantly expressed genes are also subject to the same systematic technical variations as any other spot on the chip.
Control Genes

Controls are chosen genes that have no relation to the experimental samples under study. Some Affymetrix chips employ such genes. One such type of control is known as spiked controls. They are synthetically generated DNA sequences or DNA sequences from a known organism different from the one under study. These genes are spotted on the array of the controlled chip and the variable chip. These samples have an equal amount of mRNA and thus should have equal intensities and could be used for normalization. Another controls approach is known as the titration series approach which involves the spotting of several different concentrations of the same genes or EST that span the range of intensities on the array. Unfortunately, this technique is technically challenging [29].

Scaling and Transformation

The next logical step after selecting what genes to use as references, is to scale the rest of the data to match the relative distribution of these reference genes. There are several ways to approach this. For example, one simple approach is to scale to some arbitrary constant. This constant is often a calculated value such the mean, trimmed mean or median of all the intensity values across all the experiments. Sometimes replicates would be generated to evaluate the relationships between chips and often they reveal a non-linear relationship among probes. Thus, often logarithmic transformations are performed to stabilize such variances. Recently, Lu's studies have shown that eliminating outliers did not reduce variation among multiple arrays but rather increased variation. Thus, he highly recommended using the mean of the logarithm transformed signals to calculate the normalization constant factor [34].
Data Analysis

Recall that the last step in microarray analysis is simply the statistical tests applied to the cleaned up data. Often researchers apply the common statistical procedures such as ratios and absolute differences between intensities to find differentially expressed genes. Sometimes they will use more sophisticated tests such as the popular t-test. In addition, it is quite popular to cluster the data (unsupervised or supervised) in order to find biologically relevant groups of genes. These groups of genes could be used to identify subgroups of diseases at the genetic level.

2.2.2 SAGE Technology

Like DNA microarray technology, Serial Analysis of Gene Expression was developed to provide a comprehensive and quantitative gene expression profile of target cells. The SAGE technique was developed at John Hopkins University in the USA around 1995 by Velculescu et al. [5]. Unlike the microarray technique, which are based on relative expression levels (or absolute fluorescent intensity levels (Affymetrix)), it measures mRNA expression by counting representative short sequenced tags. The expression values are thus given in absolute terms. This avoids the need for the kind of normalization required for comparing multiple microarray chips. Another important advantage of the SAGE technique is its ability to detect small transcripts, which may be crucial genes as these could be the switches or pathway regulators leading to the onset of a disease like cancer. A third advantage of SAGE over microarray is that it does not require the genes to be previously known. Thus, analysis can reveal previously unknown genes as the potential markers. Lastly, SAGE libraries may be pooled. That is, even after an experiment is done, one can go back and sequence the same tissue library even deeper and simply add these
results to the old data.

As great as SAGE seems to be, it has one major pitfall. Collecting SAGE data involves sequencing, which is a laborious and expensive process. At the current market rate, generating a SAGE library cost at least 100 times more than performing a microarray experiment. Thus, there are only a small number of SAGE libraries (a library corresponds to the measured mRNA expression levels of a sample or patient.) For instance, as will be described in Chapter 3, when we conducted the brain and breast SAGE analysis 2 years ago, there were only 17 brain cancer libraries, 7 brain normal libraries, 10 breast cancer libraries and 5 breast normal libraries publicly available through the NCBI website. Thus, while SAGE libraries give high quality data, the small sample size problem is critical for the analysis of SAGE libraries. In section 2.3, we will describe a statistical tests designed for such a problem. However, before we get into the analysis, we describe the SAGE technique in detail below.

Methodology

Recall that the basic units that make up DNA and RNA are called nucleotides. To conduct SAGE, a method called sequencing is required to read these nucleotide sequences. This procedure is complex and demands cost and time. In fact, if one were to sequence every mRNA molecule in a cell to its entirety, it would probably take a decade or so. Fortunately, as it turns out, fourteen nucleotides of an mRNA molecule is sufficient to capture the majority of the mRNA in a cell precisely. This was a major discovery as an mRNA molecule could have thousands of nucleotides. The SAGE method is based on this idea of taking a small fragment (tag) of nucleotides from an mRNA molecule to represent the entire transcript of a particular gene. These tags are obtained by special cleaving enzymes (restriction enzymes) that
recognize a specific sequence of nucleotides in the mRNA strand and cut around 14 nucleotides downstream from this recognized region. This process will almost always obtain the same sequence of tags if the transcript is the same. Of course, such a process raises the following question: *Is it possible that different genes can have the same tags?* Yes, but because there are 14 base pairs, the likelihood is statistically acceptable. However, it is likely that more than one tag represents the same gene due to biological phenomena such as polymorphism and alternative splicing. In Chapter 3, we will describe these issues in more detail and how they are circumvented. A second question that may arise is: *Is it possible that some genes may not have this restriction enzyme site?* Yes, but it is a relatively small percentage of genes as restriction enzymes typically cleave sites of no more than 4 nucleotides. Figure 2.1 briefly outlines the steps of a typical SAGE experiment (based on the figure at http://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Expression/exp82.html). Further details can be found in Appendix B.

### 2.2.3 SAGE versus DNA Microarrays

The choice of the SAGE technique over the DNA microarray hybridization technique depends on several factors including the amount of starting material, the number of samples, and the availability of resources [6]. For example, one major consideration for SAGE is the fact that around \(1.5 \times 10^6\) bases are required to be sequenced in order to do a simple 2 sample library comparison. In spite of technological advancements, sequencing is still an expensive procedure as it requires an expensive automated sequencer. Thus, when the amount of sample material is not a consideration, the DNA microarray hybridization technique is a much more cost effective technique and would yield much more results. However, if we only have a few rare precious
Figure 2.1: Experimental steps in Serial Analysis of Gene Expression (SAGE)
samples, the amplification step (PCR) in the SAGE method allows experiments to be done with sample sizes as low as 9 oocytes to 100,000 cells [6]. In fact, SAGE on a single cell has even be reported [7]. If there are a large number of samples, microarrays would be more appropriate as the costs are much lower.

2.3 Statistical Tests

Since few results in science are absolute, we must have a way to “weigh” our results. Statistics was developed to help “weigh” our results so we can determine whether an experiment is credible or not. In general, a hypothesis is generated and either accepted or rejected based on a mathematical measurement of confidence. A typical statistical experiment evaluates a population of specific values between two samples where each member has a value. For example, gene expression values between a set of cancer cell lines and a set of normal cell lines may be compared to determine how likely that this gene in question is different in cancer over normal.

2.3.1 Parametric Tests

Conventional statistics, also known as parametric tests, work well if the following assumptions are met:

1. The underlying population can be assumed. For example, a commonly used assumption is that the experimental values of the sample are normally distributed. This assumption is often referred as the parametric assumption.

2. Overall, the samples have equal variances.

3. Samples are drawn from the population independently.
If a parametric assumption cannot be met, statisticians often apply various tactics to manipulate the data. As long as this is done with care and never used to adjust for a better p-value, such a procedure is permitted. Such tactics range from deletion of outliers, winsorization, and/or trimming the data. These procedures are tricky though, as we can never be sure if these outlying values are from naturally occurring biological phenomena, or a by chance value, or a mistake in the data collection (ie. sequencing errors). Furthermore, when sample sizes are small, deleting or manipulating any values can have drastic consequences to the final results. For instance, suppose we have two samples. Sample A consists of the following values: 4, 100 and 6. Sample B consists of 5, 5 and 5. The mean of sample A is 36.7 while the mean of sample B is 5. Thus, these two samples would be deemed different by a typical statistical test. However, if the outlier 100 from sample A is removed we suddenly, sample A and B suddenly have the same mean.

2.3.2 Nonparametric Tests

Nonparametric test are classes of statistical tests that require both the equal variance and the independence assumption as described above. The advantage of these types of tests is that they do not require a parametric assumption. The disadvantage is that distribution free methods require great care because they often produce results with less efficiency in that they produce results with poor detection of false positives. However, for situations where the distribution cannot be assumed (ie. a small sample size), nonparametric tests are ideal.
Permutation Test

Often, we see inappropriate use of parametric tests conducted on only a handful of microarray or SAGE experiments. Here, we introduce the permutation test. This non-parametric test is designed for small sample sizes that makes no assumption about the underlying population. Furthermore, from figure 2.2, we can see that for every additional sample, the power of the test increases exponentially [47]. In the following Chapter, we will describe how we used the permutation test to conduct our analyses.

![Figure 2.2](image)

Figure 2.2: Illustrates that every additional sample increases the power of the test exponentially.
Cancer currently ranks second among diseases that kill humans. With the aging population and the increasing introduction of environmental complexities (ie. pollution) it is likely that it will soon overtake cardiovascular disease as the number one killer of humans. The recent development of high-throughput technologies and the recent sequencing of the entire human genome have brought great promise to unravel the mysteries to diseases such as cancer. However, progress in cancer research has been hindered by two main factors. First, cancer is a highly variable disease that is triggered by multiple genetic changes making it difficult to pinpoint genetic targets of interests. Second, current high-throughput technologies such as DNA-microarrays are not 100% reliable and often gives rise to false positives (ie. [44], [45]). Thus, the majority of findings from high-throughput technology cannot be considered ready and safe for clinical trials. For this to happen, validation of candidate genes is required which is often laborious and unsuccessful. Consequently,
it is critical that the differentially gene expression selection process is as accurate as possible.

In this Chapter, we demonstrate the importance of using an appropriate statistical test in identifying genes that are differentially expressed in tumors. As discussed in Chapter 1, one of the major problems that plagues experiments is the sample size problem. Much of the past research relied on parametric tests such as the \textit{t-test} which require a large sample size. Here, we introduce a non-parametric test known as the permutation test. This test requires a much smaller sample size as with each additional sample, the power of the test increases exponentially.

To show the appropriateness of this test we conducted two different experiments both of which apply a permutation test algorithm to identify genes that are the most differentially expressed. The first experiment has two main objectives. First, to compare the genetic signatures of two vastly different cancers (brain and breast cancer). Second, to evaluate the performance of the permutation framework over the standard \textit{t-test}. The second experiment we conducted focussed on comparing the genetic signatures of different stages of the same type of cancer; in this case lung cancer. The goal here was to find genes that were specific to each stage and genes that were common to both stages. Before getting into the details of each of these experiments, we will first describe how we preprocessed the SAGE data in each case. Following this description, we will also describe the finer details of the permutation framework we used to conduct our experiments. But before going into the details of our experiments, we will briefly discuss the recent findings in Cancer research using high-throughput technologies.
3.1 Related Work

In the past decade, the development of high-throughput gene expression technology has created a stir of enthusiasm in the cancer research community. However, as with any new technologies, often such excitement and eagerness can lead to false discoveries as many problems could be overlooked. In the following sections, we will describe some of these experiments some of which appear promising and some that appear questionable.

Applications of High-throughput Gene Expression Technologies in Brain and Breast Cancer Research

One of the main goals of cancer research is to find reliable diagnostic or prognostic biomarkers for different grades of tissue-specific tumors. Today, we rely on specialists to assess the grade and predict the behavior of the tumor based on an examination of the histology and morphology of the tumor. These methods are highly subjective and thus, is not reliable. In brain cancer for example, it has been shown that even among experienced neuropathologists, the final diagnosis is observer-dependant [46]. Thus the need for such biomarkers is desperately needed. A second goal of genomics in cancer research is to find the switches or regulators that may have caused the onset of the disease. Following such a discovery, researchers have the potential to design drugs to control these checkpoints of the pathway of the disease and thus, treat the disease.

The advent of genomic technology has provided valuable tools for cancer researchers to aid both biomarker and switch/regulator discovery. Typically, normal and cancer tissues are compared in order to identify genes that are differentially expressed. Several research groups are already demonstrating the power and poten-
tial of such technologies. However, as will be seen below, accuracy is hindered by sample sizes and noise.

Ljubimova et al. used DNA microarray technology to identify 11 genes that were deemed differentially expressed in human gliomas using a commercially produced DNA microarray called Gene Discovery Array (GDA) from Incyte Genomics, Inc. These 11 genes were chosen based on the highest mean ratios of differential expression between normal brain tissues and brain tumor tissues. The problem with such an approach is that means are sensitive to outlying values. For instance, if one microarray experiment shows a very high gene expression level due to some noise, it will be deemed significant. Not surprisingly, only 2 of these 11 genes were verified via semi-quantitative RT-PCR and Northern blot analysis [45]. These results demonstrate that the chances of false-positives are great. In another study, Watson et al. identified 196 genes that demonstrated common differential patterns among different tumor grades of oligodendrogliomas. This study successfully showed that these expression patterns could be correlated with oligodendroglioma tumor grades. The main problem with this study, is that the oligonucleotide microarray consisted of only 1100 genes and they only had 7 different samples/microarrays [46].

The accuracy of high-throughput gene expression studies for breast cancer is no better. Recently Shen et al. demonstrated the loss of annexin A1 expression in human breast cancer detected by using both SAGE and microarrays. For their microarray experiments, the used 2 normal and 7 breast cancer samples. They found 129 genes with a greater than five-fold change. To narrow the number of genes for further study, they used SAGE and EST analysis and only identified 4 qualified genes. Their expression pattern was further validated via RT-PCR. Following this analysis, ANXA1 was further confirmed at the mRNA level by Human Breast Cancer
Tissue Profiling Array and at the protein level. Only at this point, were they satisfied that this gene is of significance [42]. This demonstrates that genomic technology is far from clinical use for screening patients.

Currently, there are few studies that compare brain and breast cancer at the molecular level. Ng et al though did show that by using a clustering algorithm called OPTICS on a subset of genes selected using the Wilcoxon test, they were able to create 8 distinct clusters including brain cancer, breast cancer, brain normal, and breast normal clusters. These results suggest that brain and breast cancers are also different at the molecular level. However, no subtypes of cancers were noticeably clustered but this may have been due to the few libraries they had to work with [50].

To date, the success rate of these genome wide expression studies have not been as high as anticipated only a decade ago. Such experiments generate large lists of differentially expressed genes but only a small portion of these genes are validated upon further analysis. Because genomics data is subject to many non-biological errors, statistical significance needs to be very strong. Another common problem is small sample sizes. Undertaking a genomics experiment is a risky and expensive endeavor. In Section 3.4, we will demonstrate how much an analyses can differ just by changing how we identify significant genes. But first, we will describe how we preprocessed the SAGE data and describe the permutation framework we used to identify differentially expressed genes.

### 3.2 Preprocessing SAGE Data

Recall that SAGE has various deficiencies. First, SAGE tags often map to more than one gene due to the short sequence that represents the mRNA. Second, genes
may also map to more than one tag due to two major biological phenomena: alternative splicing and polymorphism. Both will be discussed later in greater detail. Furthermore, due to the time and money spent, the publicly available SAGE libraries created between different research institutes and universities vary in size, ranging from 10,000 to 60,000 tag types per library. In order to conduct appropriate analyses, the data must be preprocessed. Below we describe what preprocessing steps we took to minimize these deficiencies.

1. Gene to Tag Assignment

An average mature mRNA sequence is around 1800 to 2000 nucleotides in length. SAGE relies on a 10 base pair sequence to uniquely identify a gene. However, since there are only 4 different bases in total, the chances of a tag mapping to more than one gene is likely. Not surprisingly, in practice, we found that there are numerous tags that map to more than one gene. Since it is impossible to determine which gene is actually being expressed, we dealt with this problem by assigning the expression level of the tag to each gene that it mapped to. For example, if tag A maps to genes 1, 2, and 3, all the genes will be assigned the tag count of tag A.

2. Tag to Gene Collapsing

This preprocessing step deals with the inherent fact that genes often map to more than one tag. As mentioned above, this is primarily due to two major biological phenomena. The first, known as alternative splicing, describes the post processing steps of an mRNA molecule before it enters the cytoplasm. Before mRNA leaves the nucleus to be transcribed, certain regions called introns are spliced out and the remaining regions (exons) are concatenated forming a
much shorter mRNA sequence. These regions are mixed, creating alternatively spliced forms of the same gene and increases protein diversity. However, these alternate forms may or may not possess the same functionality [2] [4]. This phenomenon has recently been estimated to occur in about 55% of human genome [3]. The other major reason for multiple tags mapping to the same gene is that polymorphism among the population may exist. Polymorphism and may further contribute to several tags mapping to the same gene.

For this situation, we simply summed up the tag expression levels that mapped to the same gene. For instance, if tag A, B, and C all mapped to gene X, the expression levels of tags A, B, and C would be summed to one expression level. By performing this procedure we also get the added benefit of a reduction in the dimensionality of the data which in turn, significantly decreases computation time.

It should be noted though that this process was not done in all our experiments. If the type of experiment involves searching for only highly expressed tags, then we found it reasonable to collapse since high counts is less prone to a sequencing error. In addition, the larger counts tend to have better confidence for tag-to-gene mapping. We will discuss this issue further in Chapter 4 where collapsing is performed.

3. Scaling the SAGE libraries

To deal with the varying sizes of SAGE libraries, we scaled the tag counts to an arbitrary large total tag count (e.g., 1 million) to reduce comparison errors.
3.3 The Permutation Framework

The permutation test itself, is not a new concept or method. In fact it was introduced in the 1930s. The advent of computers have made this resampling method relatively easy to implement. Several researchers have adopted the permutation concept. For example, Dudoit et al. used the permutation test’s resampling techniques to eliminate the t-statistic’s strong normality assumption [49].

Contrary to Dudoit et al.'s work, the framework of the permutation test that we used is based strictly on a test of means between two different distributions. Here, we assume one distribution is for the gene expression level of a particular gene in normal tissues (ie. subscript n) and the other for cancerous tissue (ie. subscript c).

To select significant genes, we use the following hypotheses:

Null Hypothesis :

$H: \mu_c - \mu_n = 0$ (3.1)

Alternative Hypothesis :

$H_a: \mu_c - \mu_n \neq 0$ (3.2)

The null hypothesis states that the distribution of average gene expression levels in cancerous libraries is the same as that in non-cancerous libraries. If the
null hypothesis is rejected, it indicates that the gene expression levels of normal and
cancer samples are sufficiently different (the alternative hypothesis). A variation
of this test was previously been used for identifying differentially expressed genes
in time-course microarray experiments [48]. In the following section, we show our
specific implementation of the test. Let $N$ and $C$ be the number of normal tissue
samples and the number of cancerous tissue samples respectively.

Algorithm

1. For each gene, take all the gene counts of the normal libraries and all the
cancer libraries. Mix them up in an urn.

2. Randomly select $N$ counts from the urn to create a simulated normal set, and
calculate the simulated normal mean $\mu_{sn}$.

3. Similarly the remaining $C$ counts form the simulated cancerous set. Calculate
the *simulated* cancer mean $\mu_{sc}$.

4. Consider the random variable $v = \mu_{sc} - \mu_{sn}$, called the *simulated difference*.

5. Repeat the steps A to D above $M$ times. Let $\mu$ and $\sigma$ denote the mean and
the standard deviation of $v$.

6. Now separate the libraries back into their true identity: normal or cancerous.
Calculate the true observed difference $O = \mu_{rc} - \mu_{rn}$, where $\mu_{rc}$ denote the
true mean count of the cancerous libraries, and $\mu_{rn}$ denote the true mean of
the normal libraries.

7. Calculate the Permutation Score $PS$ where $PS = \frac{|O - \mu|}{\sigma}$.
These permutation scores statistically rate how different the two groups are from each other. The higher the permutation score is, the greater the statistical confidence that the two populations are different. While the lower the permutation score, the more statistically similar the two populations are.

In the algorithm shown above, we used $M$ permutations or iterations for each gene hit ratio. The following figure plots the hit ratio with varying values of $M$. For the results reported here, we used $M = 5,000$ on the breast data where $N = 5$ samples and $C = 10$ samples. A natural question to ask is whether this is sufficient. To answer this question, we re-ran using $M = 10,000$ on the same data and select all the genes whose permutation score exceeds a certain threshold. For example, the threshold 0.96 corresponds to the tail ends beyond 95% of the area under the standard normal curve. This set of genes is then intersected with the corresponding set of genes found using $M = 5,000$. The ratio of the size of this intersection and the size of the set of genes using $M = 10,000$ is called the hit ratio. Figure 3.1, shows that the set of genes stabilizes rather quickly, and $M = 5,000$ is a reasonable value.

### 3.4 Breast Versus Brain Cancer SAGE Libraries

To perform this experiment we used a set of brain and breast SAGE libraries found on NCBI’s public database. At the time of analysis, there were 17 brain cancer libraries, 7 brain normal libraries, 10 breast cancer libraries and 5 breast libraries.

To satisfy the first objective, we first applied the permutation test to select for the most significant differentially expressed genes in brain cancer (ie., brain cancer versus normal brain libraries). We then sorted the permutation scores in descending order and examined the top 30 ranked genes. The higher the score or rank, the more statistically confident we are that the two groups are different.
Figure 3.1: Illustrates that the higher the number of iterations, the more stable the list becomes.
Similarly, we identified the top 30 ranked genes for breast cancer. To compare the two lists, we performed an intersection of the top 500 ranked genes.

Thirty genes was chosen for two main reasons. First, each list generated over 500 genes. Looking up 500 different genes in the literature would be a very laborious task. Second, we believed that 30 genes would be sufficient to show which of the two tests gave more accurate results.

Next, to evaluate the effectiveness of the permutation test, we performed a similar experiment using the 2-sample t-test of unequal variance. To obtain the t-value, we first calculated the mean ($\mu_c$) and the pooled variance ($Sp^2$) separately for both the cancerous libraries and the normal libraries. Then we computed the t-value using:

$$ t = \frac{\mu_c - \mu_n}{\sqrt{\frac{Sp^2}{N_c} + \frac{Sp^2}{N_n}}} $$

(3.3)

The t-values were then sorted in descending order and the top 30 ranked genes were identified. This list is then compared with the permutation lists.

To evaluate the validity of our results using the two different tests, we performed a literature search on PubMed on the top ranked genes. With these top ranked genes, we looked at whether the genes were related to the neoplastic process. In order to be consistent, we used the following rules for a gene to be considered related to the neoplastic process:

1. The gene is up-regulated or down-regulated in breast or brain cancer.

2. The gene is up/down-regulated in another type of cancer.

3. The gene is a known cancer-related gene (ie. oncogene, mutator, tumour suppressor).
4. The gene is a major component of the cell cycle.

3.4.1 Summary of Validation of Breast Cancer and Brain Cancer Genes

<table>
<thead>
<tr>
<th>Category</th>
<th>Breast Cancer Related</th>
<th>Related To Other Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T-test</strong></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Permutation Test</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 3.1: Statistical Comparison of the Top 30 Genes for Breast Data

From the table 3.4.1 we can see that among the 30 top ranked genes of the breast libraries identified by the permutation test, 11 are related to breast cancer. Another 7 genes are related to other types of cancer. In contrast, among the top 30 ranked genes identified by the t-test, only 3 are known to be breast cancer related. Interestingly, all three are included in the top 30 selected by the permutation test. These results suggest that the permutation test may be more appropriate for differential gene selection when the sample size is small. For details of the exact genes found please refer to Appendix C.

<table>
<thead>
<tr>
<th>Category</th>
<th>Brain Cancer Related</th>
<th>Related To Other Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T-test</strong></td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Permutation Test</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 3.2: Statistical Comparison of the Top 30 Genes for Brain Data

Among the top 30 genes of the brain libraries identified by the permutation test, 5 are known to be related to brain cancer, and another 7 known to be related to other cancer. In contrast, among the top 30 identified by the t-test, 7 is known to be related to brain cancer, and another 8 is known to be related to other cancers.
These results show that the permutation test performed comparably with the \textit{t-test}.

The top ranked gene in the permutation test is an EST (expressed sequence tag), which is an unknown transcript that has been identified but not studied yet. The next highest ranked gene is the protein kinase C and casein kinase substrate in neurons 1. This gene is known to be highly expressed in the normal brain [43], which agrees with our results. To summarize our verification data, we have created another summary table similar to all the verified permutation score genes for our brain data experiment. For further information on the top ranked genes, we have provided complete tables of the genes found using the permutation test in Appendix C.

\subsection{Discussion}

An interesting trend to note for the literature verification data is that there seems to be a higher probability of a gene to be verified at higher ranks. For example, for the breast data, in the top 10 ranked genes we have 7 genes verified. While in the ranks 11 to 20, we have 6 genes verified. Further down, between ranks 21 and 30 we have 5 genes being verified. A similar trend was also observed in the brain data (Ranks 1-10: 6 verified, Ranks 11-20: 5 verified, Ranks 21-30: 2 verified).

Compared to the breast SAGE data results, the brain SAGE data results do not look as impressive. However, it should be noted that there are many different types of brain tissues, and thus different types of brain tumors, including gliomas, meningioma, pituitary tumours, haemangioblastoma, acoustic neuroma, pineal gland tumors, spinal cord tumors, and lymphoma. At the time of our research, we combined all the brain cancer libraries and overlooked the issue of differing brain tumor types. This is because there were not enough libraries if we were to
conduct the analysis on these tissues separately. These different types of tumors are all unlikely to have similar gene expression profiles, and thus may have contributed to the weaker results compared to the breast data.

There were also a couple other factors that may have contributed to the quality of both experiments. First, the usage of Cell Line SAGE libraries may have contributed to some inconsistencies in the data. Cell lines are tissue cells grown artificially in a petri dish and may not be representative of a true genome wide profile of a tissue taken directly from a diseased tissue (known as bulk tissues). Second, many of the top ranked genes were hypothetical proteins or unknown genes (i.e. 7 genes were hypothetical proteins for breast data) and so literature for these genes is non-existent. In addition to evaluating the breast and brain data separately, we also compared breast cancer and brain cancer at the sub-cellular level by intersecting the top 500 ranked genes to determine its similarity. A high intersection would suggest a high similarity while a low intersection would suggest a low similarity. For the top 500 permutation test results, 26/500 of the genes intersected. These results suggest a low similarity between the two cancers even at the sub-cellular level.

In spite of the few sources of error explained above, our overall results (especially for the breast data) demonstrate that the application of the permutation test on SAGE libraries has significant merit and thus, should also be used on different tissues with SAGE libraries keeping the above issues in mind. This performance could be attributed to the fact that the permutation framework is robust to outliers. The *t-test* on the other hand, would pick out spikes (outliers) in the data. For instance, suppose there were 5 libraries in each sample. For one gene, the genetic signature could look like the following: $11115$ and $11111$. The permutation test would assign this group a relatively low score and thus a low rank where as a *t-test*
would give this example a relatively higher score. Genes with no confirmed literature verification should also be further investigated via biological experimentation (ie. RT-PCR).

3.5 Biomarkers in Early Stages of Lung Cancer

3.5.1 Background and Related Work

Lung cancer is the leading cause of cancer deaths worldwide with 160,000 annual deaths in the United States alone [23]. One major reason for this is that lung cancer is very difficult to detect. In recent years, many advancements in screening technologies such as chest radiographs, computed tomography (CT) scanning, and sputum cytology have given doctor's tools for detection of cancers. However, these advancements have proven to be feeble when it comes to detecting lung cancer early enough. In fact, studies that screen high risk individuals have not been effective in decreasing mortality rates where the 5-yr patient survival rate at the clinical stage II-IV is poor ranging from 40% down to 5% [52], [51].

Today, there is a wealth of information on the histological and molecular characteristics of the premalignant changes in bronchial mucosa [52]. The earliest change that can be observed is known as reserve cell hyperplasia and squamous metaplasia. These changes spontaneously reverse upon the cessation of smoking and thus, are not considered to be true premalignant lesions. The true early premalignant changes are believed to be the low and high grade dysplasia and carcinoma in situ (CIS). Currently, it is unclear whether low grade dysplasia will lead to the advanced stages of lung cancer. However, both high grade dysplasia and CIS have been shown in various studies to lead to the invasive stages of lung cancer [53], [54], [41].
At either of these two stages, it is unlikely that these lesions will regress with the cessation of smoking.

The recent advent of global gene expression technologies have given researchers the tools to quickly generate candidate biomarkers. As far as we know, there have been no studies using such technology for evaluating stage differences at the genetic level. One reason for this is that it is very difficult to obtain a clean sample of early lesions and these lesions are rare. In the following section we demonstrate an example of how such technologies can be used for this purpose.

3.5.2 Materials

To perform our analysis, we used Lung SAGE libraries provided by BC Cancer Agency which include the following SAGE libraries:

1. 15 smoked-damaged libraries but otherwise non-cancerous (hereafter we refer to these libraries as normal). These libraries were obtained via bronchial brushings. Brushings is a medical procedure often performed during a bronchoscopy where cells from the tissues lining the respiratory tract are obtained by a small brush-like device.

2. A metaplasia library that is obtained via a bronchial biopsy. This library contains the genetic profile of the bronchial epithelium that has undergone changes due to activities such as smoking but are not cancerous. Specifically, the transformation of the normal ciliated columnar epithelium by a squamous epithelium.

3. 5 carcinoma insitu (hereafter referred as CIS) libraries obtained from a lung biopsy. CIS is the stage following severe dysplasia pre-invasive lung carcinoma
that has no metastatic potential.

4. 6 invasive libraries that were obtained via frozen resected samples. The invasive stage describes when a CIS lesion breaks through the basement membrane and has metastatic potential.

3.5.3 Candidate CIS Biomarkers

The primary objective of our analysis is to isolate genes specific to the CIS stage. In order to analyze their expression patterns we plotted a series of boxplots using the statistical and graphical software package R (http://www.r-project.org/). The figure below shows the basic parts of the boxplot. The box itself, corresponds to the interquartile range (IQR) meaning that it contains 50% of the data. The upper edge (hinge) of the box indicates the 75th percentile (median of 75% of the data) of the data set while the lower hinge indicates the 25th percentile (median of the 25% of the data). The outer whiskers represent the max and min of the data points unless an outlier exists. If there is an outlier, the whiskers are no more than the 75th percentile plus 1.5 times the interquartile range (the length of the box) or no more than 25th percentile minus 1.5 times the interquartile range (the bottom whisker). An outlier is depicted by a •.

To find these CIS specific genes we apply the permutation framework to the following:

1. CIS and Normal SAGE libraries
2. CIS and Invasive SAGE libraries

From the permutation test, we obtain a score based on the permutation score for each gene/tag that measures the statistical significance of the difference between
Figure 3.2: Diagram of a typical box and whisker plot
the two groups. Recall that the higher the score, the more confident we are that the two distributions are different. However, it should be noted that this score does not distinguish whether group A is greater than group B or visa versa. After obtaining the scores for each gene/tag, we sort them in descending order. We then took the top 2000 most differentially regulated genes of the 2 sets and performed an intersection. From this intersection, we obtained a total of 79 tags.

Figure 3.3 illustrates the steps of our analysis in a graphical manner.

Figure 3.3: Graphical Representation of Steps of Analysis for CIS Specific "Genes

To search for the up-regulated genes in CIS, but down regulated in Smoked-damaged normal tissues and invasive tissues, we first selected genes that have a permutation score greater than 2.35 for both sets (corresponding to a p-value of around 0.01). All these genes then had to meet the following requirements to be deemed a CIS biomarker candidate:

1. \( \frac{CIS}{Normal} > 10 \)
2. \( \frac{CIS}{Metaplasia} > 2 \) and
For the CIS biomarkers set, after applying the above filters we obtained 18 tags. These tags were then mapped to their corresponding unigene ids and gene names. According to the tag/gene database found at CGAP (http://cgap.nci.nih.gov/SAGE), our results mapped to 14 unigene ids and \( \frac{9}{14} \) of these mapped to genes with names. These genes included those involved in cellular adhesion, EGFR trafficking and tyrosine or phosphatase kinase receptors. Figure 3.4 illustrates the gene expression patterns specific for high expression in the CIS stage.

To search for the down-regulated genes in CIS, but upregulated in both smoked-damaged normal tissues and invasive tissues, we had to relax our requirements to get a decent output. The only requirement was that the average of CIS must be less than the average of both normal and invasive. This output generated 13 candidates. Figure 3.5 depicts the relative expression specific for low expression in the CIS stage.

Out of these 13 candidates, all mapped to unigene ids. However, 4 out of 13 of them are hypothetical and unknown genes. The remaining genes mapped to known genes. These genes included protein kinase regulators and inhibitors, and genes belonging to the annexin family. Literature shows that these genes may have potential roles in cellular signal transduction, inflammation, growth and differentiation.

### 3.5.4 Desmosomes as Candidate CIS Biomarkers

Our previous results analyzed the gene expression patterns specific for up regulation of genes in the CIS stage. Two of the genes, DSC2 and DSG3 are components of a larger adhesion structure known as desmosomes, we decided to analyze the gene
Figure 3.4: Boxplots showing gene expression patterns of genes with high CIS expression
Figure 3.5: CIS genes that are down-regulated
expression patterns of the other components (which may have been missed because of our restrictive filters) to help strengthen our observations.

Background

Desmosomes are specialized structures responsible for cellular adhesion. These highly organized intercellular junctions provide mechanical integrity to tissues by anchoring intermediate filaments to sites of strong adhesion [20]. They are made up of at least 3 major families of proteins; desmosomal cadherins, armadillo proteins and the plakins [21][22]. The cadherins (as seen in figure 3.6) are single-pass transmembrane glycoproteins that are responsible for the actual physical adhesion between two cells and as well facilitate Calcium-dependent cell to cell adhesion. These cadherins can be subdivided into the subfamilies desmogleins and the desmocollins. The desmogleins consists of 4 different types termed DSG1,DSG2,DSG3, and DSG4. Similarly, the desmocollins consists of 3 types termed DSC1,DSC2 and DSC3. Armadillo proteins include the plakoglobin, the plakophilins (PKP1, PKP2,PKP3), and p0071 [20]. Lastly, the plakins include plectin, desmoplakins I and II, and the cell envelope proteins envoplakin and periplakin [20].

As seen from figure 3.6, these groups of molecules interact in an organized fashion in order to facilitate both cell adhesion and cell rigidity/structure. The single-pass transmembrane proteins DSC and DSG are responsible for the physical adhesion between two cells. Its extracellular domains span into the dense midline (DM) where they interact with other cadherins from other cells. As seen from figure 3.6, these proteins associate closely with the tails of the desmosomal cadherins. At the same time, they also interact with plakins (desmoplakins seen in diagram). Plakoglobins and PKPs interact with the desmoplakin N-terminus to aid the cluster-
Figure 3.6: A Model of the Basic Structure of Desmosomes
ing of the cadherins and work together to form the desmosomal plaque. And finally, desmoplakin C-terminus anchors to the intermediate filaments in the inner density plaque (IDP). Proper anchoring to the IF is crucial for strengthening adhesion and tissue integrity [20].

Desmosomes and Cancer

As discussed above, desmosomes play an important role in tissue integrity and in cellular adhesion. In many types of cancers, the desmosome gene expression levels are disrupted. For example, in head, neck, and oral cancer, desmosome expression is down regulated suggesting a mechanism for a cancer to become invasive and metastasize. Such results suggest that desmosomes may have a tumor suppression functions [19]. However, this is not always the case, as in colon cancer, there is no observed downregulation prior to invasion. But this may mean that this type of cancer uses a different mechanism to invade (ie. temporary loss of adhesion in the desmosomes by other means at the protein level) [19] [16]. These results have compelled us to take a closer look at the expression patterns of all desmosome components.

Expression Patterns of Desmosome Components

To analyze the expression patterns we also used boxplots as described in Section 3.5.3.

Desmosomal Cadherins

Recall that desmoglein 3 is one of the genes that survived our strict filters. As shown in figure 3.7, the box plot clearly shows a distinct pattern where the CIS libraries
have in general, a much higher expression than the normal libraries and still higher than the invasive libraries.

![Expression pattern DSG3](image)

Figure 3.7: Expression pattern DSG3 in normal(N), metaplasia (M), CIS (C), and invasive (I) stages

Similarly, as seen in figure 3.8 the other class of cadherins, desmocollins, also show a similar gene expression pattern although the patterns are not as clear cut especially with DSC3.

**Armadillo Proteins**

The armadillo proteins are responsible for regulating desmosomal assembly and adhesion. Studies have shown that knocking out these genes led to the loss of adhesion and tissue fragility [17]. From figure 3.9, we see that DSC2 is distinctly overexpressed in the CIS stage relative to both normal and invasive stages. Note
Figure 3.8: Expression pattern of DSC2 and DSC3 in normal(N), metaplasia(M), CIS(C) and invasive(I) stages

Figure 3.9: Expression patterns of armadillo proteins in normal(N), metaplasia(M), CIS(C), and invasive(I)
that DSC2 is one of the two desmosomal cadherins extracted from our strict filters. It is also interesting to see that DSC3 also shows relative higher expression in CIS than both invasive and normal even though the difference is not as clear cut as it is for DSC2.

Plakins

The plakins are also an important part of the desmosome structure as they are critical components for proper anchorage of the intermediate filaments (IF) to desmosomes, which in turn are important for strengthening adhesion and tissue integrity [20][18]. Recall from figure 3.6 that the tail of the desmplakins are located in the inner density plaque and directly interact with both the armadillo proteins and the intermediate filaments. As seen in figure 3.10, we observe the expected gene expression patterns as other desmosome components we analyzed.

3.5.5 Invasive Candidate Biomarkers

The next part of our analysis was to analyze the gene expression patterns specific to the invasive stages. This study is similar to finding genes specific to the CIS stage. Like the CIS study, we also performed an intersection of the top 2000 most differentially expressed genes based on the permutation score with the groups CIS vs Invasive and Normal vs Invasive. Out of this intersection, we obtained a total of 78 tags. From these 78 tags, we applied the following filters:

1. \[ \frac{\text{invasive}}{\text{normal}} > 10 \]

2. \[ \frac{\text{invasive}}{\text{CIS}} > 5 \]

3. \[ \frac{\text{invasive}}{\text{metaplasia}} > 1 \]
Figure 3.10: Gene expression patterns of a desmplakin (plakin family) in normal(N), metaplasia(M), CIS(C), and invasive(I)
After applying all these similar filters, we retrieved only 6 tags. All tags have unigene ids and all mapped to genes with names. These genes included SPARC, MMP11, COL4A1, MCAM, COL1A2 and C1R. All of have unigene ids and gene symbols. Four out of the six are components of the adhesive extracellular matrix. As seen in figure 3.11, all exhibit the distinct pattern we filtered for. That is, in normal tissues, there is low or no expression while CIS has higher expression and finally invasive stages have even higher expression.

![Gene Expression Patterns](image)

Figure 3.11: Gene Expression Patterns of Candidate Invasive Biomarkers
It should be noted that some of these genes found are consistent with the literature. For example, SPARC is a secreted protein (acidic, cysteine-rich) that is a matrix-associated protein that can change shape [27]. Such a function is something one would think is important in the invasive stages. Another interesting gene is melanoma adhesion molecule (MCAM) which is a cell surface glycoprotein. According to northern blot and immunohistochemical experiments performed by Sers et al. [8], MCAM antigens are found primarily in advanced primary and metastatic melanomas. In addition to being associated with tumour progression, MCAM expression has also been found to be associated with shorter patient survival in adenocarcinoma patients. As well, it is expressed in a high portion of NSLC (non-small cell lung cancer) cells [8].

Like the CIS gene expression study, we also examined the mirror image where genes are turned off or down regulated in the invasive stages. To do this, we took the same list and applied the following filters:

1. $\frac{\text{normal}}{\text{invasive}} > 10$

2. $\frac{\text{CIS}}{\text{invasive}} > 3$

3. Fold change between invasive and metaplasia is ignored.

These filters generated 6 tags and 5 with gene names, while one tag is currently unmapped. Only one of these genes proved to be interesting. PTPRN2 (protein tyrosine phosphatase, receptor type, N polypeptide 2) is a known gene that is involved in signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. Again, to illustrate the gene patterns we plotted out boxplots at each stage as seen in figure 3.12.
Figure 3.12: Gene Expression Patterns of Candidate Invasive Biomarkers
3.5.6 Other Observed Gene Expression Patterns

To further our analysis, we also analyzed whether there are genes that are not present in CIS and Invasive stages. To find these genes, we simply performed the permutation test on the normal lung SAGE libraries and the CIS and invasive libraries pooled together. We then sorted the list in descending order based on the permutation scores and took the top 200 (all those with permutation scores > 6). Finally we select for genes with a fold change between normal and the cis/invasive pool to be greater than 40. Out of this filter, we get 15 tags with 10 genes with symbols. Figure 3.13 illustrates the selected patterns.

![Figure 3.13: Boxplots showing genes that are down regulated in lung cancer regardless of stage](image-url)
In addition to finding genes present in our normal lung samples but not in lung cancer samples, we also looked at the reverse. Here we selected for a fold change between the cancer average and normal average to be greater than 10. Performing that filter generated 6 candidates. Again, we plotted a set of boxplots to illustrate this pattern as seen in figure 3.14. At first glance, it does not appear to be very clear that the boxplots show that these genes are CIS/Invasive specific. This is because in order to retrieve these genes, we compared the mean ratios. Means are sensitive to outliers and most of the plots in 3.14 contain outliers which would skew the selected mean ratios. Thus, care must be taken when selecting these genes.

3.5.7 Summary

In summary, we have analyzed all possible expression patterns between normal, CIS and invasive stages. Specifically, we have looked at genes that are CIS specific, Invasive specific and those that are specific to both stages (CIS and Invasive). For CIS, we observed clear cut CIS specific patterns for both down and up regulation. Similarly, we observed clear cut Invasive specific patterns for both down and up regulation. However, the CIS and Invasive specific patterns were not clear cut suggesting that at the molecular level, these stages are quite different.
Figure 3.14: Boxplots showing gene expression pattern for CIS and Invasive specific stages.
Reference genes are commonly used in many gene expression experiments to attempt to normalize mRNA levels between different samples. For example, QRT-PCR and certain normalization methods in DNA microarrays both require the use of such reference genes. In the case of DNA microarrays, the normalization step attempts to isolate variations that exists due to biological differences and not due to experimental conditions. Such non-biological errors include mismatch hybridization, unequal starting material, scanner problems and spot noises. Reference genes are often used to help calibrate the intensity data to account for these errors as the expression level of these reference genes should be the same across multiple samples. However, if these reference genes are incorrect choices, then this type of normalization could severely skew the the normalized data and may even inadvertently remove important biological differences between samples especially for the more sensitive small expression values. Thus, selection of stable and abundant reference genes is a
critical step as failure to do so, could severely jeopardize subsequent analyses.

In this Chapter, we argue that reference genes must be chosen in a tissue specific manner. Furthermore, we propose a novel methodology for identification and evaluation of good reference genes. This methodology uses the advantages of the SAGE method and the permutation framework to find a stable and highly expressed set of genes (hereafter referred as the constancy requirement). In the following section, we will first describe some related works. We will then describe exactly why we use the SAGE method followed by the details of our novel methodology. Following this section, we describe how we evaluated these sets of genes and how our results suggests the importance tissue specific reference genes.

4.1 Related Work

Reference genes are typically chosen because they are biologically required and thus, usually are referred as housekeeping genes. The expression level of these housekeeping genes are assumed to be stable and highly expressed (the constancy requirement). The problem is that in the past few years, it has been increasingly apparent that many standard housekeeping genes may not be appropriate for many gene expression experiments [35],[36],[37],[40].

Recent studies by Barber et al. have shown that one of the most common housekeeping genes, Glyceraldehyde-3-phosphate (GAPDH), show variable expression. Their experiment consisted of 72 different pathologically normal human tissue types. They performed 371,088 multiplexed Q-RTPCR experiments and found significant differences in expression levels of GAPDH mRNA between tissue types and between donors of the same tissue type. Most notably, a 15-fold difference in GAPDH mRNA copy numbers was observed between the highest (skeletal mus-
cle) and lowest expressing tissue types (breast) [39]. As will be seen later, their conclusion is consistent with our results.

Zhang et al. recently reported that 3 of the 10 standard housekeeping genes were inappropriate for normalizing mRNA in neutrophils because of very low expression levels [38] and recommended only using 5 of them that showed relative stable expression. Like Barber et al.’s study, Zhang et al.’s study suggest that many standard housekeeping genes are inappropriate for mRNA normalization. The main problem with both these studies though is that they only analyze the expression level of the standard housekeeping genes which only make up less than one percent of the human genome. Thus, it is quite likely that a better reference genes exist.

Gabrielsson et al (2005) attempted to solve the reference gene problem by using 52 microarray expression profiles of human adipose tissue where they selected 50 genes with the lowest coefficient of variation [26]. The problem with such an approach is that microarrays require normalization themselves. In fact, this decade-old technology is even more sensitive to normalization than RT-PCR because of the large number of sources of systematic variation (as described in Chapter 3).

4.2 Novel Methodology

In spite of the numerous normalization techniques introduced for DNA microarrays in the past decade, to date, there is still no standard procedure. On the other hand, the SAGE (Serial Analysis of Gene Expression) technique is relatively standardized. Being a sequenced based method, SAGE technology does not require such normalization. The only preprocessing technique that is required is somewhat similar to global normalization where all samples are scaled up to some given constant. This is done to correct for unequal number of tags produced by each library as it is dif-
ficult to obtain libraries with exactly the same number of tags. In the case of DNA microarrays, the data depends on a natural hybridization reaction inherent in DNA. Furthermore, the obtained data is intensity-based and requires that we assume that the relative difference in intensity between sample A and B is approximately equal to the relative transcription between sample A and B. Arguably, this makes microarrays more susceptible to non-biological errors than SAGE. Thus, here we introduce a novel methodology to identify and evaluate reference genes based on using SAGE data.

While the SAGE methodology makes up a large part of our methodology, we also require one more component. In order to identify and evaluate potential reference genes, we use the parametric framework introduced in Chapter 3. Recall that in the permutation framework was used to identify differentially expressed genes. As it turns out, this framework can also be used to find statistically similar gene expression profiles simply by reversing the list where low permutation scores correspond to similar samples. Thus, to select for reference genes that satisfy the constancy requirement we came up with threshold that would generate a decent number of genes (20) but is constant and highly expressed. All our selected genes must meet the following criteria:

1. The permutation score must be less than 0.15
2. The average raw gene count must be at least 25 and the raw count for each library must be at least 3.
3. The standard deviation \( \sigma \) must be reasonably small (i.e., less than 10 for the lung data).

The second condition is set to guarantee that the reference gene selected
is abundant and present to some degree in all libraries while the last condition ensures that the simulated cancer distribution and simulated normal distribution are sufficiently close.

The above conditions were all applied to a set of lung SAGE libraries provided by the BC Cancer Agency. Their cohort consists of 18 non-cancerous smoke-damaged bronchial epithelial libraries, 5 pre-invasive lung carcinoma libraries (stage known as Carcinoma In situ [CIS]), and 6 invasive lung carcinoma libraries. The non-cancerous libraries are all extracted via lung brushings. Pre-invasive libraries were extracted surgically via lung biopsies and the invasive libraries were obtained from frozen resected samples.

After application of the above methodology to the lung SAGE data we obtained a list of 16 candidate reference genes. To visualize how these sets of reference genes performed, we plot the points on a scatter plot as seen in figure 4.1. The x-axis gives the permutation score indicating variability, whereas the y-axis shows the average raw SAGE counts. These sets of lung-SAGE reference genes all satisfy the constancy requirement. That is, they appear to be highly expressed and show almost constant expression across all 29 libraries of cancerous and non-cancerous libraries. Conversely, all the common housekeeping genes we examined do not meet our reference gene requirement. In fact, they are far from meeting our constancy requirement. Standard housekeeping genes such as GAPDH and TFRC in particular have a permutation greater than 3. A permutation score of 3 is equivalent to saying that the probability of the average expression in cancerous libraries differing from the average expression in non-cancerous libraries is only 1% (i.e. a p-value of 0.01).

In the next section, we describe two ways that these candidate reference genes were validated (hereafter referred as lung-SAGE reference genes).
Figure 4.1: Raw Count versus Variability on Lung SAGE data
4.3 Validation

To validate these lung-SAGE reference genes, a handful of them were analyzed by the Ontario Cancer Institute with their own cohort of lung tissues. They analyzed 7 of our lung-SAGE reference genes and 4 commonly used housekeeping genes and performed QRT-PCR experiments. QRT-PCR values were used to evaluate whether a gene satisfies the constancy requirement. If the values between tumor and normal tissues are close, then we say that they satisfy the constancy requirement. Thus, to analyze these values, we combined the two samples to compute a single standard deviation of expression. A coefficient of variation close to zero suggest that the expression level does not change much regardless of a cell’s neoplastic state. A summary of the results can be seen from figure 4.2. This figure clearly shows that in general, the standard housekeeping genes (top 4 bar lines) exhibit more variation than the lung-SAGE reference genes.

In addition to the above biological validation, we also evaluated the effectiveness of our lung-SAGE reference genes by analyzing a dataset generated by Bhattacharjee et al [33] in 2001. Essentially, we compared the same dataset normalized in two distinct manners. The first way is based on data normalized by standard housekeeping genes, while the second way is data normalized by the lung-SAGE reference genes. For both cases, we applied the same permutation framework as our previous experiment on brain and breast data to identify the top 40 most differentially expressed genes. Below we outline the exact steps we took to renormalize Bhattacharjee et al’s data:

1. Re-normalized Bhattacharjee et al’s published microarray data by fitting a line (LMS) through the SAGE selected Housekeeping genes. (Array vs Average of
Figure 4.2: Table showing relative comparison of validation results.
All Arrays).

2. Performed a permutation test on the renormalized 17 normal lung arrays vs 23 Lung Cancer Arrays

3. Performed Permutation test on Originally Normalized Lung Arrays with same arrays as above.

4. Sorted each in descending order of permutation scores.

5. Intersected the top 40 genes from SAGE Normalized and Originally Normalized Harvard Data

6. Looked up SAGE Normalized Specific Genes, Originally Normalized Specific genes and Intersected Genes and categorized them as: Up/down-regulated in Lung Cancer, Up/down-regulated in other cancers, Not previously Associated with Cancer.

Of the top 40 genes, 22 of them were in both lists. As for the other remaining genes, we categorized them into one of two cancer-related categories. The first category, as seen in table 4.1 are genes that are up/down-regulated in lung cancer while the second category are genes that are up/down-regulated in other cancer types. Our results, as seen in the table 4.1, suggest that the data normalized by lung-SAGE reference genes give results that are more consistent with the literature.

4.4 Importance of Tissue Specificity

Our second major objective is to analyze whether the choice of reference genes ought to be tissue specific. That is, a set of reference genes that satisfy the constancy
requirement for one tissue type may not be the same as the reference genes for another tissue type. To show this, we identify the sets of reference genes that satisfy the reference gene requirement for two additional tissue types, breast and brain. We use the same libraries as the ones used in our differential experiment in Chapter 3 with the exception that we only include bulk tissues (resected from patients). We took this step to keep consistent with the lung libraries produced by BC Cancer Agency which were all bulk type. Below we outline the exact steps of our analysis:

1. Performed Permutation test on Breast SAGE Libraries from NCBI which includes 12 cancer libraries and 3 normal libraries.

2. Performed Permutation test on 6 brain cancer SAGE libraries from NCBI which includes 6 normal libraries and 20 cancer libraries.

3. Performed permutation on all the above libraries including our lung SAGE libraries

4. Sorted each list in descending order to get lists of the most statistically differentially expressed genes.

After obtaining the breast and brain SAGE based reference genes, we compared them with the lung-SAGE genes. Figure 4.1 shows how breast and brain
SAGE reference genes compare with lung-SAGE genes on lung data. Notice that for both the breast and brain SAGE data, most of the genes have low average counts and relatively higher permutation scores than the lung-SAGE genes. This figure supports our theory that tissue specificity is an important factor when selecting reference genes. These findings compelled us to analyze the issue further so we looked at both the brain and breast SAGE data cases. Figure 4.3 shows a summary of our results in 3 boxplots. On y-axis we have the permutation score and on the x-axis we have the type of reference gene. In the first boxplot, we show how the permutation scores of various reference genes on lung data. Lung-SAGE reference genes show almost constant expression while standard housekeeping genes vary greatly. Interestingly, the other reference genes do not perform as badly as the housekeeping genes. The other 2 plots also show similar behavior but the variability of the housekeeping genes is comparable to the other reference genes.

Our results collectively suggest the following:

1. That standard housekeeping genes may not be appropriate for normalization of microarrays and perhaps even in other biological experiments requiring normalization (ie. QRT-PCR)

2. That there are far better reference genes than the existing ones today and tissue specific reference genes may be more appropriate.
Figure 4.3: Boxplots that depict the importance of tissue specificity
Chapter 5

Conclusion and Future Directions

5.1 Conclusion

This thesis deals with two key problems in genome wide expression analysis. The first is the small sample size problem. The second is, the normalization issue with biological experiments. Specifically, the issues surrounding poorly chosen reference genes. We have proposed using SAGE and the permutation framework to alleviate these issues. Our results in 3 independent experiments suggest that there is merit in our proposed solution.

Our first experiment compared the commonly used t-test of unequal variance and the permutation test. To evaluate their performance, each test was applied to two sets of SAGE data (brain and breast) to identify the top 30 most differentially expressed genes between normal and cancer groups. Overall, our results show that the permutation test produce results more consistent with the literature. For
instance, for the breast SAGE data, out of the top 30 ranked genes for the permutation test, 60% of the genes were verified via literature while only around 23% of the highly ranked t-test genes were verified. As for the brain SAGE data, the permutation test performed comparably with the t-test with 40% and 50% being verified respectively. However, it should be stressed that the data for the brain data was less uniform with various tumor types and may have had effect on the overall results.

The second experiment in our study further strengthened the value of the permutation test as an identifier of interesting genes. Here we analyzed lung SAGE data of various stages and performed various intersections to produce genes specific to certain stages. By performing various filters, we were able to isolate various genes specific to its stage including those specific to lung normal, CIS and invasive stages. The key findings were a couple genes known as DSG3 and DSC2. As described in Chapter 3, these two genes belong to a family of adhesion molecules known as desmosomal cadherins. The expression pattern we observe suggest that adhesion increases in pre-invasive lung carcinoma followed by a down regulation in invasive stages of the disease. Biologically, it makes sense as the definition of invasion is the spread of the cancer. For that to happen, cancer cells must lose some adhesion so that it can travel to other parts of the body. Upon this knowledge we also looked up the gene expression patterns of other associated structural genes that did not make the strict filters of our analysis. Interestingly, several of these associated genes exhibited similar gene expression patterns (but not as strong as DSG3 and DSC2). These results are now under investigation as candidate biomarkers of pre-invasive lung squamous carcinoma.

In our third experiment, we used the SAGE and permutation framework to
evaluate and identify reference genes. SAGE was used over DNA microarray data as it requires normalization itself while SAGE is immune to the normalization issues that plague DNA chip technology. In addition, recall that SAGE is a genome-wide expression technique and thus, is capable of analyzing the entire set of expressed genes in any given sample. Thus, this unique combination would be ideal for identifying novel reference genes.

5.2 The Future of Gene Expression Analysis

Gene expression analysis technologies are powerful and are showing many signs of advancing medical research. For instance, in 2002, van't Veer et al. demonstrated a strategy involving analyzing the genetic profile of individuals to help decide what breast cancer patients would benefit from certain treatments [41]. Such knowledge would be critical to a doctor's decision on what treatment to administer to the patient.

The excitement and pace of these technologies though have generated various oversights. Fortunately, in recent years, a few scientists have realized some of these oversights. For example, as discussed in the previous chapter researchers have now realized the deficiencies and limitations of using standard housekeeping genes to normalize mRNA levels [35], [36], [37],[40], [38], [39]. However, most of these studies lack the search for novel reference genes and the ones that do [26], require normalization to reference genes. Our novel methodology globally identifies genes that are the most stable and highly expressed (what we called the constancy requirement). We showed that the genes it selected can be verified via QRT-PCR and showed that normalizing DNA microarray data with these newly acquired reference genes produce results more consistent with the literature. Thus, future experiments
must consider methods such as ours, to conduct normalization on mRNA levels.

There is still much to be done beyond our study. In fact, so far we only grazed the surface of this SAGE data that we analyzed. This data is very large and there are just numerous combinations of ways to look at the data. Recall from Chapter 3 that we generated many different figures showing many different gene expression patterns. All such patterns still need to be analyzed carefully to pull out any interesting genes that map to known pathways or structures. Following this analysis, we would analyze the gene expression patterns of the genes in the pathway which can involve several genes. In addition to these studies, we also could look at low expressors. Low expressors may be interesting to study because they may be the switches or regulators that cause the onset of the disease. The good news is that this study could also be done using the permutation framework that we have outlined because during our analysis, we found that the permutation test also pulls out results with low counts.

After extracting genes that belong to some existing pathways and examining its gene expression pattern within our own SAGE data, we still need to conduct some sort of validation. Recall from Chapter 3 on the second experiment with the lung SAGE data, our objective was to find candidate biomarkers for early stages of lung cancer. Following the analysis, we selected two of these candidates for analysis: namely DSG3 and DSC2. In spite of our efforts to show that there is strong evidence from the SAGE data that desmosome components are up-regulated just before the invasive stages of lung cancer, we cannot take what we observe and publish it. Thus, often, researchers perform an extra step of analysis; namely biological validation. This is due to three reasons. First, biological validation can show that our observations are also true at the protein level (ie. via antibody staining). Second, this
extra step conducted on a different cohort can make our findings more statistically convincing. Third, genome wide expression technologies are subject to more errors than other small scale technologies. For instance, in SAGE, one thing that is constantly changing is the tag-to-gene mapping database. Throughout our study, we had to update our results various times because of this issue, which sometimes led to a small but significant change in the results. Thus, validation can help distinguish artifact from true observation. Unfortunately, biological validation can take several months to a year. Furthermore, since our analysis attempts to extract CIS specific genes, simply obtaining additional CIS samples for biological validation would take even more time.

We were very fortunate to have collaborated with experts in the biological domain. As you recall, some of our experimental results were validated by the Ontario Cancer Institute. Specifically, the housekeeping project, as described in Chapter 4. Their biological experiment was performed on a different cohort and helped validate and strengthen our findings. Without the wet-lab and dry-lab relationship, our research could only go so far. Our results are strong in both the wet-lab and dry-lab side. Thus, strongly suggesting that housekeeping genes do not make the best reference genes and that normalizing to reference genes is appropriate if we select them the way we have.
Bibliography


Appendix A

Central Dogma of Biology

The smallest biological unit of life is known as a cell. Some organisms are unicellular while others are multicellular. Unicellular organisms contain all the machinery to sustain life and reproduce. Multicellular organisms are much more complex as they require interactions between other cells to obtain necessary nutrients to sustain life and also often require interactions between other multicellular organisms to reproduce. Different cells produce different gene products. The activation of specific gene products or proteins can have a great effect on the characteristics of a cell as they are the actual worker molecules of the cell. For instance, a muscle cell and a nerve cell express different gene products and thus, differ in both function and physical appearance.

The central dogma of biology explains just how a gene product is produced in cells. Cells contain genetic templates known as DNA located in a region called the nucleus. These gene templates are copied into corresponding molecules known as messenger RNA (mRNA) through a process known as transcription. mRNA molecules (or often called transcripts) are then exported into the cytoplasm where
they are recruited by structures called ribosomes (rRNA). These structures in turn recruit free amino acids (the building blocks of proteins) and perform a process known as translation where the ribosome interpret the instructions contained in the transcript until finally the entire protein is created.

DNA/RNA consists of subunits known as nucleotides (A Adenine, T Thymine (DNA)/ U Uracil (RNA), C Cytosine, G- Guanine). DNA in its natural state consists of 2 strands of DNA where each nucleotide have an affinity for another corresponding nucleotide and forms a robust double stranded structure known as the double helix. Specifically, adenine has an affinity for thymine and cytosine has an affinity for guanine and vice versa. This natural phenomena is known as hybridization and as you will see later, is a key reaction for DNA microarrays to work. RNA on the other hand, are single stranded and do not have thymine but rather a similar molecule known as uracil.
Appendix B

Genome-wide Expression Technologies

B.1 DNA Microarrays

A simple microarray experiment requires 6 basic steps:

1. Sample preparation
2. Array fabrication
3. Hybridization of the sample to the array
4. Scanning and image analysis
5. Normalization/Data Preprocessing
6. Data Analysis

Sample preparation

Sample preparation is considered to be the most sensitive step in a microarray experiment step. One reason for this is that samples must be extracted from crude...
sources such as blood, tumor biopsies, and complex multi-cellular tissues. Thus, purity of samples becomes a major issue. For instance, a scientist maybe only interested in the tumour tissue but in order to extract a tumor, healthy tissues are also extracted along with it. In addition, such samples must be treated with chemicals such as formalin to preserve the tissue which may alter or degrade the RNA and thus, make it unusable. Fortunately, this is an active area of research where new technologies have been providing scientists with tools to help extract purer samples. For instance, in recent years, the development of new technologies such as laser capture microdissection has allowed scientist to extract purer samples. LCM (laser capture dissection) can help eliminate unwanted cells and create a purer sample [15].

Array fabrication

The primary concerns when manufacturing a DNA microarray is precision, cost and time. Today, DNA chips are either manufactured in the lab (especially in academia) or ordered from various specialized companies that market and manufacture them including Affymetrix, Amersham Pharmacia Biotech, Biorobtoics, and Genomic Solutions. The manufacturing process is also a hot area of research as scientists and businesses aim to reduce cost and time without loss of precision or to increase precision with little cost. For instance, one of the first technologies developed, photolithographic DNA synthesis, allowed the placement of high-density oligonucleotides on microarray slides. However, the procedure is costly and time consuming, and most importantly, the precision is poor as the synthesized oligonucleotides are subject to a wide variation and uncertainty. Thus, alternatives were developed namely mechanical microspotting, ink jet ejection and bubble jet technology which helped alleviate
various pitfalls of older technologies [13].

Hybridization of the sample to the array

After the cell samples are captured (either grown or extracted from bulk tissues directly) they are placed in a test-tube. These test-tubes are centrifuged in order to move the cells to the bottom of the test-tube and then treated with an RNA extraction compound. This compound is then extracted and placed in another test tube and reversed transcribed to produce cDNA molecules as DNA molecules are much more stable than RNA molecules. As the cDNA is made, it is tagged with a fluorescent dye. These fluorescently labeled probes are then placed on a DNA microarray chip. DNA microarray chips contain many spots of complementary bases and are either made up of oligonucleotides (synthetically produced short sequences) or entire cDNA sequences. For spotted microarrays, two samples (the reference and the experimental) are labeled with different dyes (ie. Cy3 and Cy5) and competitively hybridize onto the chip. As mentioned above, for Affymetrix type of arrays, controls and experimental samples are hybridized on separate microarray slides and thus produce arrays of absolute expression intensities.

Scanning and image analysis (ie. Spot location, background correction, intensity assignments)

After hybridization, fluorescent dyes cannot be seen until a specialized laser passes over the slide. While this is done, light is emitted and captured by a computer and saved as an image file. Over the years, various groups have developed microarray image analysis software to extract the varying intensities for each spot in an array. Up to now, there is no standard way of doing this specifically for microarrays as there
are many different image analysis methods (http://oz.berkeley.edu/tech-reports/). However, all microarray image analysis software has the same goals. First, on top of calculating the spots associated signal intensities, image analysis also aims to assess local background noises. In addition, most of these software tools also allow for basic filtering to deal with ghost spots (background higher than intensity spot), damaged spots, and very low intensity spots. Much of these poor spots are caused by experimental factors such as wobbling of the robot arm that makes the deposits of the samples onto the slide and contamination (ie. dust, thumbprints) [14].

Normalization/Data Preprocessing

Refer to Chapter 3.

Data Analysis

Refer to Chapter 3.

B.2 SAGE Technology

The SAGE methodology basically works by capturing the mRNAs in cells. Since most mRNA molecules end with a long string of As (polyadenylated tails), scientist have discovered that they can easily capture these molecules by using a long string of the contemplenary base pairs, thymine (T). To do such a procedure though, mRNA must be transcribed into cDNA (DNA version of an mRNA transcript) since DNA is a much more stable structure than RNA. This process is performed by an enzyme found in RNA viruses called Reverse Transcriptase. The next important step is to bind a molecule with 20 or more Ts to special microscopic magnetic structures called streptavidin beads. This allows the capturing of mRNA molecules in the
cytoplasm of the cells. These beads can then be withdrawn with a magnet along with the hybridized DNA molecule (a copy of the original RNA molecule). Next, as shown in the above figure, this cDNA molecule is cleaved by a restriction enzyme that creates a sticky end hanging out (CTAC). This sticky end allows linkers to be attached to these ends. Then a tagging enzyme, which detects the linker site, reaches down past the CTAC sticky end and cuts off a short segment of the cDNA molecule. This short segment is the tag and ideally contains enough nucleotides to identify the original mRNA molecule. After the enzymatic cutting steps, the ends of these short molecules have special characteristics that allow them to bind together. Together they form longer structures called concatemers. These structures are then PCR amplified so they can be detected and then fed into a sequencer that counts and adds up all unique tags.
Appendix C

Validation of Differentially Expressed Genes

C.1 Validation of Differentially Expressed Genes of Breast Data

1. Annexin A1 (UID: 782255):
   (a) Rank: 1
   (b) Score: 3.75
   (c) Description: They are thought to have a role in tumour suppression and are upregulated in many cancers including breast cancer.
   (d) References: PMIDS: 9062391, 9514092, 8387039

2. Serum Amyloid A2 (UID: 336462)
   (a) Rank: 4
   (b) Score: 3.32
(c) Description: Elevated in many types of tumours including breast cancer.

(d) References: OMIM: 104751, PMID: 6200925, 3734116, 11596022

3. Ribosomal protein S15; RPS15 (UID: 133230):

(a) Rank 5

(b) Score: 3.32

(c) Description: Found to be activated in various human tumors such as insulinomas, esophageal cancers, and colon cancers.

(d) References: OMIM: 104751, PMID: 6200925, 3734116, 11596022

4. nuclear receptor subfamily 2, group F, member 6; NR2F6 (UID: 239752):

(a) Rank: 6

(b) Score: 3.30

(c) Description: Members of this family of genes is involved in control and differentiation in neoplasia. Absent in human colon carcinoma

(d) References: OMIM: 132880, PMID: 2553781

5. interleukin 8: (UID: 624)

(a) Rank: 7

(b) Score: 3.18

(c) Description: Studies suggests a role for IL-8 in promoting the metastatic potential of breast tumor cells

(d) References: OMIM: 146930, PMID: 11330965, 11159200, 11029796

6. protein tyrosine phosphatase, non-receptor type 13 (UID: 211595):

(a) Rank: 8

(b) Score: 3.17
(c) Description: Is expressed in normal breast epithelial cells and is frequently up-regulated in breast cancer

(d) References: OMIM: 600267, PMID: 11696979, 9544992, 10640988

7. nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha: (UID: 81328)

(a) Rank: 9
(b) Score: 3.10
(c) Description: Activation of nuclear factor-kappaB in breast cancer
(d) References: OMIM: 600664, PMID: 11325857, 11752211

8. LIM domain protein (UID: 79691):

(a) Rank: 11
(b) Score: 3.09
(c) Description: Overexpressed in breast cancer.
(d) References: OMIM: 603422, PMID: 11734645

9. transforming growth factor beta-stimulated protein TSC-22 (UID: 114360):

(a) Rank: 12
(b) Score: 3.07
(c) Description: Transcription factor and belongs to the large family of early response genes and thought to have a tumour suppressor role in many cancers.
(d) References: OMIM: 607715, PMID: 11944908, 11836610, 11095965, 10879745, 10854535, 9459148, 9458104, 9195978

10. serum amyloid A1: (UID: 332053)
11. inositol polyphosphate-1-phosphatase (UID: 32309)

(a) Rank: 15
(b) Score: 3.05
(c) Description: Upregulated in human colorectal cancer
(d) References: OMIM: 147263, PMID: 10747296, 8392378

12. superoxide dismutase 2, mitochondrial (UID: 318885):

(a) Rank: 17
(b) Score: 3.01
(c) Description: Down regulated in lung carcinoma.
(d) References: OMIM: 147460, PMID: 11313974, 11491651

13. protein phosphatase 1, regulatory (inhibitor) subunit 15A (UID: 76556):

(a) Rank: 18
(b) Score: 3.01
(c) Description: Member of a group of genes whose transcript levels are increased following stressful growth arrest conditions and treatment with DNA-damaging agents. Its protein response is correlated with apoptosis following ionizing radiation.
(d) References: PMID: 11593419, 11836553

14. intercellular adhesion molecule 1 (CD54), human rhinovirus receptor (UID: 168383):
15. spermidine/spermine N1-acetyltransferase; SAT (UID: 28491):
   (a) Rank: 23
   (b) Score: 2.93
   (c) Description: Activity in both breast cancer cells and other types of cancers (ie. Lung)
   (d) References: OMIM: 313020, PMID: 8216356, 2731159, 12697027

16. baculoviral IAP repeat-containing 3 (UID: 127799):
   (a) Rank: 24
   (b) Score: 2.91
   (c) Description: Inhibit apoptosis (programmed-cell death)
   (d) References: OMIM: 601721, PMID: 8643514, 8552191

17. prostate epithelium-specific Ets transcription factor (UID: 79414):
   (a) Rank: 26
   (b) Score: 2.90
   (c) Description: The mRNA is overexpressed in human breast tumors and is a candidate breast tumor marker and a breast tumor antigen.
   (d) References: PMID: 11555586

18. keratin 14 (epidermolysis bullosa simplex, Dowling-Meara, Koebner); KRT14 (UID: 117729):
(a) Rank: 30
(b) Score: 2.86
(c) Description: Expressed in breast cancer.
(d) References: OMIM: 148066, PMID: 10786689, 11487275

C.2 Validation of Differentially Expressed Genes of Breast Data

1. protein kinase C and casein kinase substrate in neurons 1 (UID: 6462):
   (a) Rank: 2
   (b) Score: 4.26
   (c) Description: Strong expression in normal brain. Co-localize and bind with dynamin. Phosphorylates AGT which controls the susceptibility to methylate carcinogens in tumor cells.
   (d) References: OMIM: 606512, PMID: 11023825, 10667577, 9746365, 11179684, 11082044, 10667577

2. synuclein, beta, (UID: 90297):
   (a) Rank: 3
   (b) Score: 4.25
   (c) Description: Upregulation in ovarian and breast tumors.
   (d) References: OMIM: 602569, PMID: 8194594, 10048491, 10813729

3. ATPase, Ca++ transporting, plasma membrane, 2 (UID: 89512):
   (a) Rank: 4
   (b) Score: 4.21
(c) Description: Upregulated in brain.

(d) References: OMIM: 108733, PMID 10533058

4. Protein phosphatase-2A, regulatory subunit B' (PR 53), (UID: 236963):
   (a) Rank: 6
   (b) Score: 4.18
   (c) Description: Reduced expression of the Aalpha subunit of protein phosphatase 2A in human gliomas in the absence of mutations in the Aalpha and Abeta subunit genes. Thought to function as tumour suppressors.
   (d) References: OMIM: 600756, PMID: 11519040

5. Enolase-2, gamma, Neuronal, (UID: 146580):
   (a) Rank: 8
   (b) Score: 4.17
   (c) Description: Elevated in glioma cells
   (d) References: OMIM:131360, PMID: 7520111, 6268172

   (a) Rank: 10
   (b) Score: 4.12
   (c) Description: Associated with various brain tumors
   (d) References: OMIM: 602377, PMID: 11072786, 1832879, 10749171

7. Visinin-like 1, (UID: 2288):
   (a) Rank: 12
   (b) Score: 4.08
   (c) Description: Plays an important role in regulating tumor cell invasiveness and that its loss could aid in enhancing the advanced malignant phenotype
8. **UID: 75149 SH3 domain, GRB2-like, 2**

   (a) **Rank:** 14
   (b) **Score:** 4.06
   (c) **Description:** Preferentially expressed in the brain.
   (d) References: OMIM: 600817, PMID: 9364517, 12941826

9. **Prostatic binding protein (UID: 80423):**

   (a) **Rank:** 16
   (b) **Score:** 3.99
   (c) **Description:** Used to regulate the onset of mammary and prostate cancer in transgenic mice.
   (d) References: OMIM: 604465, PMID: 9169142

10. **RAS-associated protein RAB3A (UID: 27744)**

    (a) **Rank:** 19
    (b) **Score:** 3.97
    (c) **Description:** Located in the chromosome region 19p13.1-p12. 19p13.2 site is involved in malignant processes such as acute leukemias.
    (d) References: OMIM: 179490, PMID: 8432525

11. **Kinesin 2, 60-70kD (UID: 117977):**

    (a) **Rank:** 25
    (b) **Score:** 3.93
    (c) **Description:** Breast cancer antigen (upregulation)
    (d) References: OMIM: 600025, PMID: 9177777, 12747765
12. *myelin transcription factor 1-like (UID: 172619)*:

(a) *Rank*: 28

(b) *Score*: 3.85

(c) *Description*: Upregulated in high-grade human brain tumors

(d) References: OMIM: 600379, PMID: 9210873
Appendix D

Normalization of Expression Using Permutation Test and SAGE (NEPS) Reference Genes

D.1 Breast-SAGE Reference Genes

D.2 Brain-SAGE Reference Genes

D.3 Lung-SAGE Reference Genes

D.4 All-SAGE Reference Genes
<table>
<thead>
<tr>
<th>#</th>
<th>Symbol</th>
<th>Gene</th>
<th>Score</th>
<th>Stdev.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ITM2B</td>
<td>integral membrane protein 2B</td>
<td>0.14</td>
<td>9.48</td>
<td>11.73</td>
</tr>
<tr>
<td>2</td>
<td>TSPYL1</td>
<td>TSPY-like</td>
<td>0.13</td>
<td>18.12</td>
<td>10.40</td>
</tr>
<tr>
<td>3</td>
<td>CDH1</td>
<td>cadherin 1, type 1, E-cadherin (epithelial)</td>
<td>0.060</td>
<td>20.75</td>
<td>12.13</td>
</tr>
<tr>
<td>4</td>
<td>TAP</td>
<td>T-cell activation protein</td>
<td>0.05</td>
<td>16.42</td>
<td>10.13</td>
</tr>
<tr>
<td>5</td>
<td>HLA-DQB1</td>
<td>major histocompatibility complex, class II, DQ beta 1</td>
<td>0.046</td>
<td>18.87</td>
<td>10.20</td>
</tr>
<tr>
<td>6</td>
<td>CGI-135</td>
<td>CGI-135 protein</td>
<td>0.042</td>
<td>13.37</td>
<td>10.20</td>
</tr>
<tr>
<td>7</td>
<td>DKFZp564F053</td>
<td>Homo sapiens mRNA; cDNA DKFZp564F053 (from clone DKFZp564F053 )</td>
<td>0.033</td>
<td>16.54</td>
<td>10.60</td>
</tr>
<tr>
<td>8</td>
<td>PPIB</td>
<td>peptidylprolyl isomerase B (cyclophilin B )</td>
<td>0.024</td>
<td>20.48</td>
<td>22.80</td>
</tr>
<tr>
<td>9</td>
<td>ATP1B</td>
<td>ATPase, Na+/K+ transporting beta 1 polypeptide</td>
<td>0.022</td>
<td>16.86</td>
<td>10.73</td>
</tr>
<tr>
<td>10</td>
<td>GK001</td>
<td>GK001 protein</td>
<td>0.020</td>
<td>13.20</td>
<td>11.27</td>
</tr>
</tbody>
</table>

Table D.1: Breast-SAGE Reference Genes
<table>
<thead>
<tr>
<th>#</th>
<th>Symbol</th>
<th>Gene</th>
<th>Score</th>
<th>Stdev.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TCEB2</td>
<td>transcription elongation factor B (SIII), polypeptide 2 (18kDa, elongin B)</td>
<td>0.15</td>
<td>9.59</td>
<td>10.81</td>
</tr>
<tr>
<td>2</td>
<td>FLJ22678</td>
<td>hypothetical protein FLJ22678</td>
<td>0.14</td>
<td>11.16</td>
<td>13.27</td>
</tr>
<tr>
<td>3</td>
<td>ENO1</td>
<td>enolase 1, (alpha)</td>
<td>0.14</td>
<td>10.074</td>
<td>12.73</td>
</tr>
<tr>
<td>4</td>
<td>P23</td>
<td>unactive progesterone receptor, 23 kD</td>
<td>0.14</td>
<td>14.020</td>
<td>16.81</td>
</tr>
<tr>
<td>5</td>
<td>KIAA0193</td>
<td>KIAA0193 gene product</td>
<td>0.13</td>
<td>11.75</td>
<td>11.88</td>
</tr>
<tr>
<td>6</td>
<td>TTYH1</td>
<td>tweety homolog 1 (Drosophila)</td>
<td>0.12</td>
<td>14.49</td>
<td>10.81</td>
</tr>
<tr>
<td>7</td>
<td>ZYX</td>
<td>zyxin</td>
<td>0.089</td>
<td>12.39</td>
<td>10.88</td>
</tr>
<tr>
<td>8</td>
<td>DDAH1</td>
<td>dimethylarginine dimethylaminohydrolase 1</td>
<td>0.088</td>
<td>13.43</td>
<td>10.54</td>
</tr>
<tr>
<td>9</td>
<td>ATP5G2</td>
<td>ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9), isoform 2</td>
<td>0.077</td>
<td>13.99</td>
<td>12.46</td>
</tr>
<tr>
<td>10</td>
<td>NCKAP1</td>
<td>NCK-associated protein 1</td>
<td>0.049</td>
<td>13.25</td>
<td>11.12</td>
</tr>
<tr>
<td>11</td>
<td>PPP1CB</td>
<td>protein phosphatase 1, catalytic subunit, beta isoform</td>
<td>0.047</td>
<td>12.32</td>
<td>10.96</td>
</tr>
<tr>
<td>12</td>
<td>MACF1</td>
<td>microtubule-actin crosslinking factor 1</td>
<td>0.019</td>
<td>10.89</td>
<td>10.5</td>
</tr>
<tr>
<td>13</td>
<td>CCT7</td>
<td>chaperonin containing TCP1, subunit 7 (eta)</td>
<td>0.016</td>
<td>10.44</td>
<td>10.69</td>
</tr>
<tr>
<td>14</td>
<td>SAP18</td>
<td>sin3-associated polypeptide, 18kDa</td>
<td>0.0070</td>
<td>14.70</td>
<td>11.81</td>
</tr>
<tr>
<td>15</td>
<td>PFKL</td>
<td>phosphofructokinase, liver</td>
<td>0.0030</td>
<td>12.34</td>
<td>10.50</td>
</tr>
</tbody>
</table>

Table D.2: Brain-SAGE Reference Genes
<table>
<thead>
<tr>
<th>#</th>
<th>Symbol</th>
<th>Gene</th>
<th>Score</th>
<th>Stdev.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RPL11</td>
<td>ribosomal protein L11</td>
<td>0.0043</td>
<td>8.57</td>
<td>71.45</td>
</tr>
<tr>
<td>2</td>
<td>RPL15</td>
<td>ribosomal protein L15</td>
<td>0.023</td>
<td>4.94</td>
<td>44.03</td>
</tr>
<tr>
<td>3</td>
<td>RPL17</td>
<td>ribosomal protein L17</td>
<td>0.027</td>
<td>7.99</td>
<td>63.72</td>
</tr>
<tr>
<td>4</td>
<td>PKM2</td>
<td>pyruvate kinase muscle</td>
<td>0.054</td>
<td>7.51</td>
<td>66.65</td>
</tr>
<tr>
<td>5</td>
<td>MRLC2</td>
<td>myosin regulatory light chain</td>
<td>0.054</td>
<td>4.48</td>
<td>40.52</td>
</tr>
<tr>
<td>6</td>
<td>RSP13</td>
<td>ribosomal protein S13</td>
<td>0.087</td>
<td>4.44</td>
<td>30.52</td>
</tr>
<tr>
<td>7</td>
<td>ATP5A1</td>
<td>ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit isoform 1, cardiac muscle</td>
<td>0.093</td>
<td>7.58</td>
<td>81.34</td>
</tr>
<tr>
<td>8</td>
<td>ATP5J</td>
<td>ATP synthase, H+ transporting, mitochondrial F0 complex, subunit F6</td>
<td>0.097</td>
<td>2.65</td>
<td>29.48</td>
</tr>
<tr>
<td>9</td>
<td>H3F3B</td>
<td>H3 histone, family 3B (H3,3B)</td>
<td>0.097</td>
<td>2.91</td>
<td>28.34</td>
</tr>
<tr>
<td>10</td>
<td>CST3</td>
<td>cystatin C (amyloid angiopathy and cerebral hemorrhage)</td>
<td>0.099</td>
<td>4.21</td>
<td>31.48</td>
</tr>
<tr>
<td>11</td>
<td>PRDX1</td>
<td>peroxiredoxin 1</td>
<td>0.11</td>
<td>7.32</td>
<td>53.66</td>
</tr>
<tr>
<td>12</td>
<td>SRP14</td>
<td>signal recognition particle 14kDa (homologous Alu RNA binding protein)</td>
<td>0.12</td>
<td>4.00</td>
<td>42.31</td>
</tr>
<tr>
<td>13</td>
<td>NINJ1</td>
<td>ninjurin 1</td>
<td>0.12</td>
<td>3.06</td>
<td>28.14</td>
</tr>
<tr>
<td>14</td>
<td>RPL4</td>
<td>ribosomal protein L4</td>
<td>0.14</td>
<td>4.65</td>
<td>41.03</td>
</tr>
<tr>
<td>15</td>
<td>TEBP</td>
<td>unactive progesterone receptor, 23 kD</td>
<td>0.14</td>
<td>4.65</td>
<td>41.03</td>
</tr>
<tr>
<td>16</td>
<td>NDUFV1</td>
<td>NADH dehydrogenase (ubiquinone) flavoprotein 1, 51kDa</td>
<td>0.14</td>
<td>4.00</td>
<td>29.00</td>
</tr>
</tbody>
</table>

Table D.3: Lung-SAGE Reference Genes
<table>
<thead>
<tr>
<th>#</th>
<th>Symbol</th>
<th>Gene</th>
<th>Score</th>
<th>Stdev.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UQCRB</td>
<td>ubiquinol-cytochrome c reductase binding protein</td>
<td>0.15</td>
<td>1.98</td>
<td>23.30</td>
</tr>
<tr>
<td>2</td>
<td>RPL6</td>
<td>ribosomal protein L6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>G22P1</td>
<td>thyroid autoantigen 70kDa (Ku antigen)</td>
<td>0.14</td>
<td>2.14</td>
<td>21.13</td>
</tr>
<tr>
<td>4</td>
<td>FLJ20003</td>
<td>hypothetical protein FLJ20003</td>
<td>0.13</td>
<td>6.38</td>
<td>37.07</td>
</tr>
<tr>
<td>5</td>
<td>SRP9</td>
<td>signal recognition particle 9kDa</td>
<td>0.12</td>
<td>4.23</td>
<td>42.56</td>
</tr>
<tr>
<td>6</td>
<td>GTF2I</td>
<td>general transcription factor II, i</td>
<td>0.11</td>
<td>2.08</td>
<td>24.30</td>
</tr>
<tr>
<td>7</td>
<td>ARHA</td>
<td>ras homolog gene family, member A</td>
<td>0.11</td>
<td>2.62</td>
<td>31.10</td>
</tr>
<tr>
<td>8</td>
<td>VPS28</td>
<td>vacuolar protein sorting 28 (yeast)</td>
<td>0.11</td>
<td>2.070</td>
<td>24.50</td>
</tr>
<tr>
<td>9</td>
<td>SERP1</td>
<td>stress-associated endoplasmic reticulum protein 1</td>
<td>0.084</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CAPNS1</td>
<td>calpain small subunit 1</td>
<td>0.083</td>
<td>0.083</td>
<td>4.28</td>
</tr>
<tr>
<td></td>
<td>43.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>RPS6</td>
<td>ribosomal protein S6</td>
<td>0.063</td>
<td>3.81</td>
<td>35.27</td>
</tr>
<tr>
<td>12</td>
<td>RPS2</td>
<td>ribosomal protein S2</td>
<td>0.062</td>
<td>7.33</td>
<td>69.40</td>
</tr>
<tr>
<td>13</td>
<td>LAPTM4A</td>
<td>lysosomal-associated protein transmembrane 4 alpha</td>
<td>0.050</td>
<td>3.11</td>
<td>31.99</td>
</tr>
<tr>
<td>14</td>
<td>RPL3</td>
<td>ribosomal protein L3</td>
<td>0.048</td>
<td>8.69</td>
<td>101.14</td>
</tr>
</tbody>
</table>

Table D.4: All-SAGE Reference Genes