**Title**
Challenges in High-throughput Data Analysis: Proteomic Data Pre-processing and Network Methods for Integrating Multiple Data Types



**Author**
Liao, Eileen

**Publication Date**
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

# Challenges in High-throughput Data Analysis: Proteomic Data Pre-processing and Network Methods for Integrating Multiple Data Types

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Biostatistics

by

Eileen Lingchen Liao

2012

# ABSTRACT OF THE DISSERTATION

## Challenges in High-throughput Data Analysis: Proteomic Data Pre-processing and Network Methods for Integrating Multiple Data Types

by

Eileen Lingchen Liao

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2012

Professor Robert Elashoff, Chair

1) Proteomic Data Pre-processing: Quantification and Normalization of Luminex Assay System

High through-put genomic and proteomic technologies allow rapid analysis of molecular targets of thousands of genes at a time, either at the DNA, RNA or protein level. In these type of experiments variations in expression measurements can occur from a variety of sources. Our goal was to examine measurement and normalization techniques to reduce the experimental variation in data derived from a bead-based multiplex Luminex assay system which allows simultaneous measurements of proteins. Normalization for the Luminex assay system requires a fundamentally different approach than the case of traditional microarrays. In the Luminex assay system, each experimental unit is a plate and each plate has results for multiple subjects and analytes. We quantified performance among different measurement systems (fluorescent intensity, background

in fluorescent intensity, and observed concentration) in both high and standard scanning systems. Various normalization techniques (scale normalization, quantile normalization, lowess curve normalization) were adapted to the Luminex data scenario and their performance was compared in two datasets.

We used the coefficient of variation across plates to compare the performance of normalization methods. Median and Lowess normalizations appeared to result in reducing plate-to-plate variation the most. Quantile normalization does not appear to work well for these datasets. Our results suggest that simple normalizations such as scale and lowess curve normalizations perform better than complex methods such as quantile normalization. Complex methods may add noise and bias to the normalized adjustment when the assumptions are not met.

2) Integration of microRNA and mRNA by Weighted Gene Co-expression Network Analysis

We focus on the step-by-step network construction and module detection of mRNAs by weighted gene co-expression network analysis (WGCNA), followed by identifying the strong correlation between miRNA and module eigengenes. We then evaluate whether the predicted mRNA targets are differentially present between a given module and other modules by using the Fisher's exact test. We retained miRNAs who are significant in the fisher exact test, and are strongly correlated with eigengenes in a module.

Next we relate modules to disease status by using eigengene network methodology, we found that 11 out of 13 modules are significantly related with disease status. Enrichment analyses by DAVID software are implemented for the 11 modules.

We also run step-by-step network construction and module detection of miRNAs and found 6 modules. We used LASSO regression to explore the relationship between miRNA and mRNAs.

The predictors are module eigengene of miRNA and the outcome is the eigengene from each mRNA module.

We found that 1 miRNA “hsa_miR_25” is significantly anti-correlated with mRNA Magenta module. “hsa_miR_25” belongs to the miRNA module “blue” that is also predictive to Magenta mRNA module through LASSO regression. Its putative mRNA targets are found and integrated from the renal dataset.

In conclusion, the weighted co-expression network analysis provides a novel integrative view of miRNA and their putative genes. It also greatly alleviates the multiple testing problems that plague standard gene-centric methods.

The dissertation of Eileen Lingchen Liao has been approved.

Steven M. Dubinett

Christina R. Kitchen

Steve Horvath

David Elashoff

Robert M. Elashoff, Committee Chair

University of California, Los Angeles

2012

To my parents, my husband Victor, and my daughter Katelyn whom I deeply love..

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I could not have completed the four-year journey to obtain my Ph.D. without the guidance, encouragement and support of many people. First and foremost, I need to acknowledge my advisor Professor David Elashoff for his excellent mentorship. I will forever treasure the lessons he drilled into me on how to do scientific research with independent thinking. He trained me to carefully plan and justify each statistical approach so that even failed solutions can yield useful information. He also trained me with scientific writing and communication skills which are indispensable for career growth.

I also greatly appreciate of the support and guidance from Professor Robert Elashoff for his constructive suggestion on my dissertation research.

Professor Steve Horvath is kindly and resourcefully to collaborate the second project with me. I am deeply grateful that he spends much time discussing the implementation of the second project with me, provides the guidance to write the draft, and organize and submit the paper.

I would like to take this opportunity to thank Professor Steven Dubinett for kindly providing the data and supporting for my GSR.

I also want to thank Professor Christina Kitchen for kindly serving on my doctor committee and providing insightful comments in dissertation.

The four-year journey has definitely not been smooth sailing all the way. And for that, I am grateful to have the support of my family. My Mom and Dad took the initiative to help take care of my daughter Katelyn. Without their help and spiritual support, I would not have saved enough energy and time to prepare the oral exam and do research. I would not have walked to the end of

this journey. I also greatly thank my husband Victor for his encouragement, calm and humor. He never doubts my abilities to finish the Ph.D. program.

# VITA

| | |
|---|---|
| 2004 – 2006 | M.S.<br>Department of Biostatistics<br>University of California, Los Angeles |
| 2006 – 2008 | Biostatistician<br>Amgen<br>Thousand Oak, California |
| 2011 | Summer Biostatistics Interns<br>Abbott Laboratory<br>Chicago, Illinois |
| 2011 – 2012 | Teaching Assistant<br>Department of Biostatistics<br>University of California, Los Angeles |
| 2011 – 2012 | Graduate Student Researcher<br>School of Medicine<br>University of California, Los Angeles |

# PUBLICATIONS AND PRESENTATIONS

**Liao E,** Elashoff D, Horvath S**.** Integration of miRNA and mRNA by Weighted Gene Co-expression Network Analysis. *In preparation to BMC Systems Biology*.

Babbitt C, Halpern R, **Liao E**, and Lai K. Hyperglycemia is associated with intracranial injury in children less than 3 years of age. *Pediatric Emergency Care*. Accepted July, 2012.

John M, Eliezer N, **Liao E**, Inderpal R. Beta-Blocker Management of Refractory Hemoptysis in Cystic Fibrosis: A Novel Treatment Approach. *Chest*. Submitted April, 2012.

Weight S, Derhovanessian A, **Liao E**, Hu S, Gregson A.L, Kubak B.M, Saggar R, Plachevskiy V, Fishbein M.C., Lynch J.P., Ardehali A, Ross D.J, Wang H.J, Elashoff R.M., and Belperio J.A. CXCR3 Chemokines Ligands During Respiratory Viral Infections Predict Lung Allograft Dysfunction. *American Journal of Transplantation*. 2012 Feb; 12(2): 477-84.

Elashoff D**, Liao E,** Dubinett S, Weight S. Gardner B. Data Preprocessing: Quantification and Normalization for the Luminex Assay System. *In preparation*.

Fan K, Andrews B, **Liao E,** Allam K, Amaral C, Bradley J. Protection of Temporomandibular Joint Problems during Syndromic Neonatal Mandibular Distraction using Condylar Unloading. *Plastic and Reconstructive Surgery*. 2012 May; 129(5):1151-61.

Fan K, Roostaeian J, Sorice S, **Liao E**, Tabit C, Tanna N, Bradley J. Evaluation of Plastic Surgery Training Programs: Integrated/Combined vs. Independent. *Plastic and Reconstructive Surgery*. Accepted December 27, 2011.

Babbitt C, Cooper M**, Liao E**. A single Center's Experience with High-Frequency Oscillatory Ventilation in Children. *Respiratory Care*. Accepted June, 2012.

Fan K, **Liao E**, Dickinson B, Bradley J. Reply: Rank Sum Test or Paired t Test? *Plastic and Reconstructive Surgery*. 2011 Oct; 128 (4): 369-370e.

Lim A, Fan K, Allam K, Wan D, Tabit C, **Liao E**, Bradley J, Kawamoto H. Autologous Fat Transplantation in the Craniofacial Patient: UCLA Experience. *Journal of Craniofacial Surgery*. Submitted August 24, 2011.

Levin, L, Kitchen C, **Liao E**, Kim B, Marrogi A, Widney D, Krampf R, Magpantay L, Breen E.C, Martinez-Maza, O. Elevated serum levels of CXCL13 precede the diagnosis of B cell non-Hodgkin lymphoma. *Blood,* submitted, 2011.

Gregson A, Hoji A, Palchevskiy V, Hu S, Weigt S, **Liao E**, Derhovanessian A, Saggar R, Song S, Elashoff R, Yang O, and Belperio J. Protection Against Bronchiolitis Obliterans Syndrome is Associated with Allograft $CCR7^{+}CD45RA^{-}$ T Regulatory Cells. *PLoS One*. 2010 Jun 29; 5(6): e11354.

Patel KR, White SC, Tejirian T, Han SH, Russell D, Vira D, **Liao E**, Patel KB, Haigh P, Gracia C, Dutson E, Mehran A. Gallbladder management during laparoscopic roux-en-y gastric bypass surgery: routine pre-operative screening for gallstones or post-operative prophylactic medical treatment are not necessary. *The American Surgeon*. 2006 Oct;72 (10):857-61 17058721

**Eileen Liao**, David Elashoff, Steve Horvath. Integration of miRNA and mRNA by Weighted Gene Co-expression Network Analysis. Joint Statistical Meeting. Poster Presentation. August 30, 2012. San Diego.

**Eileen Liao**. Chair the High Dimensional Data Session. ENAR. April 3, 2012. Washington DC.

**Eileen Liao**, David Elashoff. Data Preprocessing: Quantification and Normalization for the Luminex Assay System. ENAR. Poster Presentation. March 31, 2012. Washington DC.

Ahmed Sulliman, Kenneth Fan, Neil Tanna, **Eileen Liao**, Jaco Festekjian. Autologous Breast Fat Grafting: A survey of Current Opinions and Practices among Plastic Surgeons. American Society of Reconstructive Microsurgery. January 14-17, 2012. Las Vegas, NV.

Ahmed Sulliman, Kenneth Fan, Neil Tanna, **Eileen Liao**, Jaco Festekjian. Autologous Breast Fat Grafting: Current Opinions and Practices among North American Plastic Surgeons. 9$^{th}$ Annual Meeting of the International Federation of Adipose Therapeutics and Sciences. November 4-6, 2011. Miami, FL.

**Eileen Liao**, David Elashoff. Data Preprocessing: Quantification and Normalization of the Luminex Aasay System. Joint Statistical Meeting. August 2, 2011. Miami, FL.

**Eileen Liao**, Vipin Arora. Application of Multiple Imputation in Immunology. Abbott Intern Poster Presentation. July 21, 2011. Abbott Park, IL.

Ahmed Suliman, Kenneth Fan, Neil Tanna, **Eileen Liao**, Malcolm A Lesavoy, Jacob Festekjian. Autologous Breast Fat Grafting – Current Opinions and Practices among North American Plastic Surgeons. UCLA Division of Plastic Surgery, Annual Resident Research. June 14, 2011. Los Angeles, CA.

# Chapter 1: Introduction

Over the last several years, there has been an increased demand for high-throughput, cost-effective and accurate measurement of small proteins and other analytes in clinical trial and research laboratories. High throughput genomic technology has become an essential tool to understand gene regulation and interaction [1]. High throughput techniques, such as cDNA microarrays[2], bead-based multiplex immunoassay[3], miRNA analysis, proteomics arrays enable the analysis of thousands of gene expressions simultaneously.

In the past few years researchers are moving from single analyte assay to multiplex bead-based solutions, such as the Luminex bead-based multiplex assay system[4]. The data are automatically processed and evaluated by the analysis system. The results are represented by multi-analyte profile and directly provide the intensity quantification. Using this process, Luminex bead-based assay allows rapid and simultaneous multiplexing of up to 500 analytes with a single sample.

In high throughput technology, there are many sources of systematic variation. Measurements may be systematically biased by diverse effects such as efficiency of RNA extraction, reverse transcription, label incorporation, exposure, scanning, spot detection, etc. Data pre-processing is a critical factor in assuring the validity and success of downstream studies. The key-processing steps are quantification and normalization[5]. In the Luminex system results are quantified before communication to the researcher, however normalization of data across plates may still need to be performed to reduce systematic sources of variation.

There are a variety of techniques to remove systematic variation. Yang YH, Dudoit S et al.[6] summarized a number of normalization methods for dual labeled microarrays such as global

normalization, intensity dependent normalization and within-print-tip-group normalization. Smyth GK et al.[7] introduced a few lowess normalization including print-tip lowess and composite lowess normalization. Amararunga[4], Quackenbush J [8] and Bilban et al.[9] gave a thorough review on normalization methods. Chen L & Wong WH [10-11]proposed model-based analysis of oligonucleotide arrays. There are some extensions for global and intensity-dependent normalization. For example, Chen et al. [12] proposed a subset normalization to adjust for location biases, and global normalization for intensity biases. Edwards[13] proposed non-linear lowess normalization to correct for spatial heterogeneity. Bolstad et al.[14] developed the Quantile normalization algorithm and showed that it performed well for microarray data.

Normalization approaches are designed to remove systematic variations, while retaining biological variation. Visual aids such as MA plots [15-17]can be used to assess the effectiveness of normalization methods in cDNA microarray. Quantitative criteria to assess normalization methods include: rank variation of spot intensity in non-normalized versus normalized data [15, 18] correlation [19-20]; variance[17, 21]; error in replicated data[16, 22].

In this thesis, we will present and attempt to address several general as well as data-specific statistical issues in high throughput technology. After all, the data themselves do not express the knowledge. They have to be first analyzed, and the associations and relationship could be studied and confirmed before conclusions are drawn.

The background information important to high throughput technologies will be presented in the first chapter, including basic understand of the field of genomics, the microarray technology and the principle of microarray. Chapter 2 describes the main preprocessing methods in microarray analysis, especially quantification and normalization of microarray. Chapter3 describes four normalization methods for a lung cancer marker dataset and contains comparative

results for each method. Chapter 4 we validated the normalizations methods for a lung transplant dataset. In chapter 5 we introduced the method to integrate miRNA and mRNA datasets.

## 1.1 Genomics Basics:

Genes, often referred to as the discrete hereditary units, are made of *deoxyribonucleic acid* (DNA). A DNA molecule consists of a sugar group, a nitrogenous base and a phosphate group. The base groups for DNA can be one of four types: adenine (A), thymine (T), guanine (G), and cytosine (C). Because of the favorable hydrogen-binding interactions, these molecules are capable of forming base pairing where adenine (A) pairs with thymine (T) and guanine (G) pairs with cytosine(C). DNA molecules can also form a polymer through the chemical bonds between the phosphate group and the sugar group, which allow them to become a long DNA strand. When two complimentary DNA strands come in contact with each other, the base pairings cause the two strands wind tightly with each other in spiral structure known as double helix[23].

The DNA sequence, a particular order of the bases arranged along one strand, encodes the information necessary for virtually all aspects of an organism's biological functions [24]. The central dogma of molecular biology formulates the transmission of genetic information from DNA to protein as: DNA → mRNA → protein. A simplified description of this process can be broken down to two stages. In the first stage, DNA is transcribed into a transient intermediary molecule known as messenger RNA (mRNA). The mRNA is similar to DNA with three key differences:1) mRNA is single-stranded, 2) its sugar group is replaced with ribose and 3) the base thymine (T) in the DNA is substituted by uracil (U). The second stage is translation, during which mRNA serves as the template for protein synthesis. In this stage, the coding regions of mRNA read three bases at a time in units known as codons and convert them into a string of amino acids that folds into a protein molecule. Thus mRNA is translated into a protein. Finally,

protein acts as building blocks and the workhorse of life, and it regulates most of life's day-to-day functions[25].

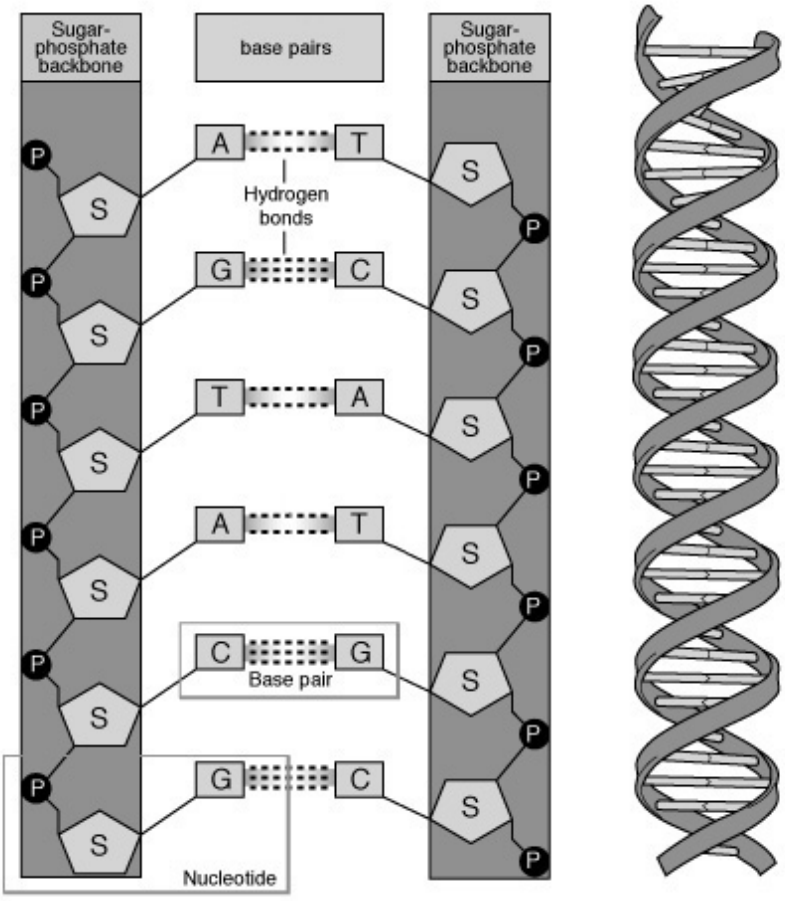


**Figure 1: DNA structure**

In biological research, although the protein molecules are useful in connecting genetic information to function, it is often inaccurate as a temporal measurement for the level of expression for the gene. This is due to a number of factors such as the longevity of the protein molecules, sensitivity of protein detection assays and problems with specificity and yields during the protein purification. Unlike proteins, mRNA levels, which are composed of intermediate molecules, are relatively simple to purify and analyzed by high-throughput technology. So to date, attention has been primarily focused on expression at the transcription stage, i.e. on mRNA levels.

The different types of microarray systems, including cDNA microarray, Affymetrix oligonucleotide array, Luminex bead-based assay, have now become the widely used technology to study mRNA, miRNA, proteomic levels. These technologies have provided a means of detecting the expression patterns of a huge amount of genes at once, thereby bring out tremendous improvements over the traditional PCR assay that can only analyze a few genes per experiment.

## 1.2 Affymetrix High-Density Oligonucleotide Microarray

GeneChip arrays are the combination of advanced technology, design criteria, and quality control processes. Affymetrix's GeneChip manufacturing is directed by photochemical synthesis. Because of this technology, more than a million different probes can be synthesized on an array that is the size of a thumbnail. Over the years, Affymetrix's platform has proven to be a reliable and robust system for the genome-wide analysis of gene expression. A typical Affymetrix microarray experiment[4] process follows the next 5 steps:

1) Preparing the microarray:

The Affymetrix array begins with a 5-in quartz wafer. Through two basic steps: deprotection and coupling, the oligonucleotide, a short chain composed of up to approximately 20 nucleotides, is synthesized on the wafers. Once synthesis is complete, wafers are diced in a variety of array sizes for use. Typically a 1.28 $cm^2$ array can accommodate more than 1.4 million different probe locations, and each of these probes contains millions of identical DNA molecules.

2) Preparing the labeled sample:

The mRNA, the working complementary copies of target genes within cells, is purified from the cells following conventional methods. In order to detect which mRNA is recognized by the

microarray, the sample is labeled by fluorescent dyes that fluoresce when exposed to the correct wavelength of light. The labeled sample is the target for the experiment.

3) Hybridizing the labeled sample to the microarray and washing the microarray:

Affymetrix utilizes two types of probes: 1) probes that are completely complementary to the target sequence. These are called perfect match probe (PM). 2) probes with a single mismatch to the target, centered in the middle of the oligonucleotide. These are mismatch probe (MM). MM are identical to the PM sequences, except differs on the 13th nucleotide. Mismatch probes are designed to account for the effects of nonspecific binding, cross-hybridization, and electronic noise. Affymetrix refers to each PM-MM pair as a probe pair. And it uses 11-20 probes, which have 25 oligonucleotide bases, to represent one gene. The entire set of probe pairs for a gene is called probe set. In each array, these probe sets are selected to interrogate specific single nucleotide polymorphisms. In addition, arrays contain a number of different control probes which are used for quality control.

The labeled sample is then sealed to a hybridization chamber to allow the hybridization reactions to complete. The single-stranded mRNA molecule binds tightly with high affinity to PM that has precisely matched sequence, and binds with significantly lower affinity to MM that has imperfect match. Due to this base pairing, the mRNA preferentially hybridizes to the oligonucleotides spot in each array.

The microarray is then washed to eliminate the excess labeled sample, and dried with clean compressed air.


**Figure 2: A probe set (PM and MM) in Affymetrix microarray**

4) Scanning the microarray:

Since the sample is labeled with fluorescent reporter molecule that emits detectable light, the microarray is scanned to determine the amount of labeled sample. Scanner not only picks up light emitted by the target mRNA, but also inevitably picks up lights from various sources, such that the labeled sample hybridized nonspecifically to the glass slide, unwashed labeled samples that is adhered to various other chemicals used in the process, etc. All these signals constitute background which needs to be accounted for.

Also, scanner sets up lower and upper threshold intensity levels for each measurement. When intensities exceed the upper threshold, saturation is occurred. So there is a trade-off between precision and threshold: increasing one will decrease the other.

5) Interpreting the scanned image

After the microarray is scanned, the intensity of fluorescence on each spot is recorded. The end product of this microarray experiment is a scanned gray scale image, which is usually stored in a proprietary file format. Various softwares could convert the image into spot intensity measurements, which can be analyzed for gene expression differences. Higher intensity

corresponds to more mRNA expression levels of a gene in the sample. A typical microarray image is shown Figure 2. Red spots are from samples that are transcribed to cDNA and labeled with red fluorescence. Green spots are from other samples that are labeled with green fluorescence. When both cDNA is hybridized to the same location, mixed signal will produce the yellow color.

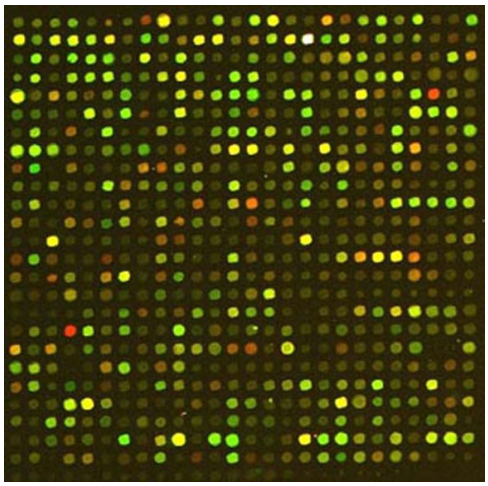

**Figure 3: A scanned microarray image**

## 1.3 Luminex Bead-Based Multiplex Immunoassay

The Luminex bead-based assay is a new promising technology. As researcher focus on multiple targets, they are moving from single analyte assays to multiplex bead-based solutions, such as the Luminex assay system. This fiber-optic system is a bundle of optical fibers where microscopic wells are etched on the end of each fiber[4]. These wells hold the DNA sequences in bead form. Each bead can be coated with a reagent specific to a particular bioassay. Once the array is exposed to the fluorescently labeled sample, and sample finds the complementary DNA sequence on the array, the hybridization takes place. A light source will illuminate the array and excite fluorescent probe in the tagged samples. This causes a signal to be passed through the optical fiber to a detector, which could identify which probe matches sequences in the labeled

sample. Using this process, Luminex bead-based assay allows rapid and simultaneous multiplexing of up to 500 analytes with a single sample.

Figure 4 details the process of multiplex assays that are directly performed in microliter plates. Test reagents contain a mixture of different microspheres populations (bead mix). Each bead type is characterized by an individual red fluorescent color tone and a capture reagent (A, B, C). These specific detection reagents are oligonucleotide probes or specific proteins staying on bead's surface.

During incubation, microsphere population (bead mix) react with a patient sample, and captures its specific target molecules (analyte). The analytes are bound to the surface of the microspheres, and could be labeled using a green fluorescent labeling reagent (marker or conjugate). The amount of bound analytes corresponds directly to the fluorescent intensity quantification.

Within a few seconds, thousand of microspheres are analyzed individually for their red (analyte classification) and green (reporter quantification) fluorescence using a Luminex assay system. The data are automatically processed and evaluated by the analysis system. The results are represented by multi-analyte profile and directly provide the intensity quantification.

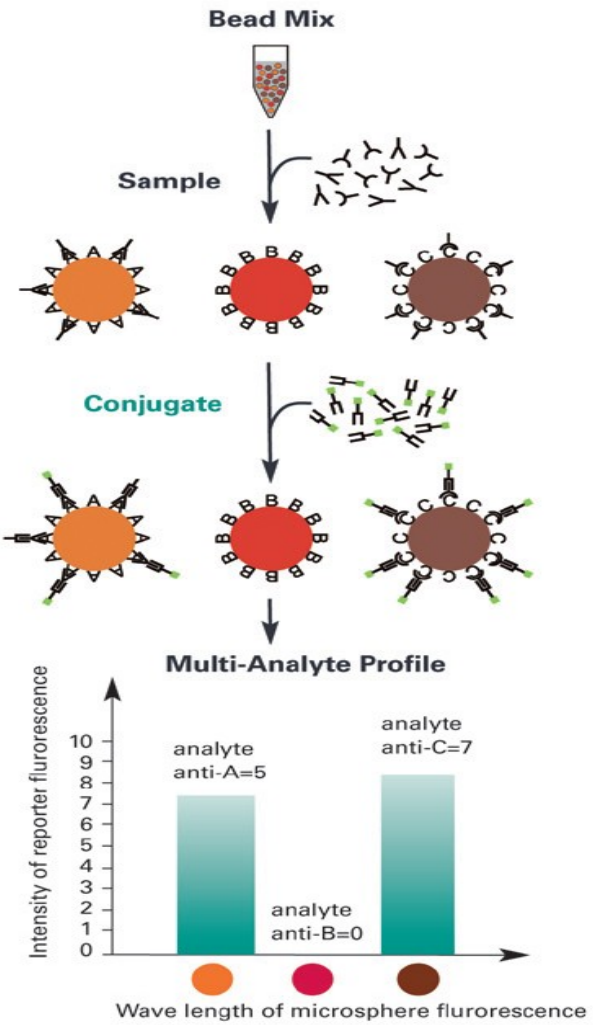


**Figure 4: Luminex multiplex detection reactions**

## 1.4 Microarray Experiment Principle

To explain the principle behind the microarray experiment, a simple hypothetical example is illustrated here.

Suppose we have obtained cancer tissue and normal tissue and we would like to know which genes are expressed differentially between two tissues. We first extract mRNA from each tissue, then we reverse-transcribe mRNA to cDNA and conjugate fluorescent dyes to each sample. The labeled samples are often referred to as the "target".

Suppose we have a DNA microarray containing 36,000 genes. Each of these 36,000 genes are printed on the rectangular array of spots on the tiny glass slide, with one gene corresponds to a spot. These genes are called “probe”. Two such microarrays are prepared.

Then we flood one microarray with labeled samples from cancer tissue, and the other microarray with samples from normal tissue. We allow enough time for cDNA in the samples to hybridize to probes in the slide. Then we wash off the excess labeled samples from the microarray and dry them.

In principle, each spot in the microarray could identify a gene that corresponds to some reverse transcribed mRNA in the labeled sample. These spots are easily recognized as they are the spots that will fluoresce. We will then scan the microarray and measure the intensity level of fluorescence. By comparing the intensities between two microarrays, we are able to tell which genes are differentially expressed between cancer and normal tissues.

# Chapter 2: Preprocessing Microarray Data

Once the microarray experiment is performed and spot intensity data are collected, these data are ready to be analyzed to draw conclusions. Transforming raw data into a scale suitable for analysis, removing systematic source of variation, and identifying outliers in array are common methods in preprocessing. Normalization is one of common the methods to pre-process microarray data.

In early microarray experiments, variation of expression measurements among arrays can be attributed to many sources, such as differences in mRNA preparation, cDNA labeling, image intensity and microarray hybridization/wash efficiency. Variation still exists despite huge improvements in the technology. These variations include interesting biological variation and unwanted non-biological variation. It is the unwanted obscuring variation, sometimes called technical variation, which normalization procedures seek to remove. Sources of obscuring variation could be differences in scanner-setting, mRNA hybridized quantities, processing order, differences in labeling efficiency between two fluorescent dyes and many other factors. Normalization attempts to remove such variation which affects the measured gene expression levels.

Yang YH, Dudoit S et al [26]summarized a number of normalization methods for dual labeled microarrays that includes global normalization and locally weighted lowess smoothing. Following this, good reviews on normalization were provided by Quackenbush[8] and Bilban[9]. The extension of normalization for global and intensity-dependent normalization was later brought up by Kepler[22] and Wang[27]. Kepler et al[22] utilized local regression to estimate a normalized intensities and intensity-dependent error variance. Wang et al[27] proposed an iterative algorithm to estimate normalized coefficients and identify control genes in cDNA

microarray. Workman et al.[15] proposed a robust non-linear method for normalization using array signal distribution analysis and cubic spline. Chen et al. [12] proposed a subset normalization to adjust for location biases combined with global normalization for intensity biases. For cDNA microarray, Edwards [13] considered a non-linear lowess normalization in one channel to correct spatial heterogeneity.

A consideration that needs to be addressed is: when applying normalization method to microarray data, how many genes are expected to change between different conditions (such as treatment and control), and how these changes will occur. Most normalization methods require that the number of genes changing in expression between different conditions is small. When this assumption is not satisfied, ideally the normalization would only be applied on arrays within each treatment condition group.

Although normalization alone cannot control all systematic variations, normalization plays in important role in the early stage of microarray data analysis because expression data can significantly vary from different normalization procedures. Subsequent analyses, such as clustering, gene networks, they are all dependent on the performance of a normalization procedure.

A few normalization methods used in cDNA microarray, Affyemetric microarray, and Luminex assay system are described in the following section. These approaches could be used separately or in combination to normalize a set of microarrays.

## 2.1 Normalization for cDNA Microarray

**1) Notation in an array display**

For a spot j, j = 1, …, p, let $R_j$ and $G_j$ denote the measured fluorescence intensities for the red and greed dyes respectively.

Denote log intensity ratio $\mathrm{M} = \log_2 R/G$, and mean log-intensity $A = \log_2 \sqrt{RG}$. An M vs. A plot is widely used representation of the (R, G) data in terms of log-intensity ratio M, which is the interest to most studies.

**2) Within-Slide: Global Normalization**

In this case, the normalization is done separately for each slide, using the red and green intensities for each slide.

Global normalization assumes that red and green intensities are proportional to each other, i.e. $\mathrm{R} = \mathrm{k} \cdot G$, so that the center of the distribution of log-ratios is shifted to be 0:

$\log_2 R/G \to \log_2 R/G - c = \log_2 R/(kG)$, where $c = \log_2 k$ is the median or mean of the log-intensity ratios for a gene set. Global normalization is still the most widely used methods in spite of intensity dependent dye biases in numerous experiments.

**3) Within-Slide: Intensity-Dependent Normalization**

In most cases, the dye bias appears to be dependent on spot intensity, which is shown in M vs. A plot, where M is the log intensity ratio and A is the mean log-intensity. The lowess smoother is used to perform a local A-dependent normalization:

$\log_2 R/G \to \log_2 R/G - c(A)$

Where $c(A)$ is the lowess fit to the M vs. A plot. The lowess() function is a scatter-plot smoother which performs robust locally linear fits. In addition, the lowess() function will not be affected by a small amount of differentially expressed genes which appear to be outliers in the M vs. A plot. This will be discussed in detail again in Affymetrix array.

**4) Within-Print-Tip-Group Normalization**

When we take into account the fact of every grid in an array, within-print-tip-group normalization is a common choice. It is simply a (print-tip + A)-dependent normalization, that is,

$\log_2 R/G \to \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G)$

Where $c_i(A)$ is the lowess fit to the M vs. A plot for the *ith* grid only, $i = 1, \cdots, I$, and $I$ represents the number of print-tips.

**5) Within-Slide Normalization: Scale**

After the within-print-tip group normalization introduced above, it is possible that the log-ratios from various print-tip have different spread and scales. One approach to adjust this is assuming all log-ratios from the *ith* print-tip group follow a normal distribution with mean 0 and variance $a_i^2\sigma^2$, where $a_i^2$ is the scale factor for the *ith* print-tip group, and $\sigma^2$ is the variance of the true log-ratios. The constraint for this is $\sum_{i=1}^{I} \log a_i^2 = 0$, where $I$ denoting the total number of print-tips in the array, so that the estimated $a_i$ is:

$\hat{a}_i = \dfrac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt{\prod_{k=1}^{I} \sum_{j=1}^{n_i} M_{kj}^2}}$ Where $M_{ij}$ is the *jth* log ratio in the *ith* print-tip group, $j = 1, \cdots, n_i$.

A more preferable robust alternative to this estimate is: $\hat{a}_i = \dfrac{MAD_i}{\sqrt{\prod_{i=1}^{I} MAD_i}}$, where MAD is the median absolute deviation: $MAD_i = median_j \{| M_{ij} - median_j(M_{ij}) |\}$.

Behind this estimation there are a few assumptions: 1) it assumes a small number of genes are expressed significantly different between two mRNA samples. 2) it assumes the spread of the distribution of the log-ratios is roughly same for all print-tip groups. In brief, the statistics MAD, is robust and is not affected by a small percentage of differentially expressed genes.

## 2.2 Normalization for Affymetrix array

### 2.2.1 Scale Normalization

There are a few variations of scale normalization. One such method is normalization by the sum. This method normalizes the sums of intensities of each microarray to be equal to one another, which is 1.

The assumption of doing this is the total RNA content is roughly the same across samples. Suppose for each microarray their sums are $x_{1+}, x_{2+}, \cdots x_{k+}$. If we divide all the observations in the *ith* array by $x_{i+}$, their sum will be 1. By doing this for all the microarrays would make the sum equal to1 for each microarray.

Another version of this method is normalization by the mean, in which the arithmetic means of the microarray are equated. Similarly, normalization by the median result in median intensities is the same across all arrays. Q3 normalization, in which the 3rd quartiles are equated, is also commonly used. The rational for using Q3 normalization is that we expect about half of the genes are unexpressed and the 3rd quartile is the median intensity of the rest expressed genes.

All these methods normalize the intensities by scaling. By comparing multiple arrays and normalizing them, we assume the overall distribution of mRNA intensity values does not change between samples, and most genes change very little in intensity across samples. These assumptions are reasonable since we start with equal quantities of mRNA for the samples we are going to prepare. Therefore the average hybridization should be same for all the samples.

Normalization by scaling belongs to the global or linear normalization schemes. The common feature of these schemes is it assumes the spot intensities on every pair of arrays being normalized are linearly correlated, so that the lack of comparability could be corrected by adjusting every spot intensity on any microarray by the same amount. This same amount of

intensity is called normalization factor. The disadvantage of scale normalization is it depends on how the baseline array is chosen. If the poor baseline is chosen, the poorer result is produced.

The scale normalization could be performed by Affymetrix MAS software in the following steps:

1. Choose a baseline array
2. For each array *i* besides baseline, multiply each probe expression value by:

(probe value)* [(mean expression on baseline array)/(mean expression on array *i*)]. This makes each array having the same mean intensity as baseline array.

Below is the example of how to perform mean normalization in each step.

Step1: calculate average intensity for each slide:

| | Slide 1 (baseline) | Slide 2 | Slide 3 |
|---|---|---|---|
| Probe 1 | 10 | 25 | 50 |
| Probe 2 | 20 | 40 | 70 |
| Probe 3 | 25 | 50 | 60 |
| Probe 4 | 15 | 45 | 40 |
| Mean | 17.50 | 40 | 55 |

Step 2: Since we choose slide 1 as baseline, then multiply each cell by average 1, then divide by average of array *i*.

| | Slide 1 (baseline) | Slide 2 | Slide 3 |
|---|---|---|---|
| Probe 1 | 10 | 25*17.5/40 | 50*17.5/55 |
| Probe 2 | 20 | 40*17.5/40 | 70*17.5/55 |
| Probe 3 | 25 | 50*17.5/40 | 60*17.5/55 |
| Probe 4 | 15 | 45*17.5/40 | 40*17.5/55 |
| Mean | 17.50 | | |

Step 3: Now average intensities of each array are all the same – average of slide 1

| | Slide 1 (baseline) | Slide 2 | Slide 3 |
|---|---|---|---|
| Probe 1 | 10 | 10.94 | 15.90 |
| Probe 2 | 20 | 17.50 | 22.27 |
| Probe 3 | 25 | 21.88 | 19.09 |
| Probe 4 | 15 | 19.69 | 12.73 |
| Mean | 17.50 | 17.50 | 17.50 |

**Table 1: Mean normalization**

### 2.2.2 Quantile Normalization

Bolstad et al[14] in 2003 introduced this method to minimize non-biological difference that might exist in multiple arrays. It assumes the distribution of intensities for each array the same.

This is motivated by Q-Q plot where the distribution of two vectors are the same if the plot is a straight diagonal line, and are not same if it is other than a diagonal line. If we extend this idea to n dimension, then plotting the quantiles in n dimension would give a straight line along the line given by the unit vector $(\frac{1}{\sqrt{\mathrm{n}}},\ldots,\frac{1}{\sqrt{n}})$. Following this idea, if the points of n dimensional quantile plots are projected onto the diagonal line, it suggests the set of data have the same distribution. This is how the quantile normalization is based on.

The algorithm for normalizing a set of vectors by quantile normalization is as follows:

1) Given n arrays of length p, form a matrix X of dimension $\mathrm{p}\times\mathrm{n}$ where each gene is a row, and each array is a column

2) Sort each column of X separately to generate a sorted $\mathrm{p}\times\mathrm{n}$ matrix Y

3) Take the mean of each row of Y and generate a p-dimensional vector $\mathrm{Y}_{sort}$

4) Set averages as value for all elements in the row

5) Rearrange vector $\mathrm{Y}_{sort}$ to have the same ordering of the corresponding column of matrix X so that the empirical distributions of intensities are the same as that of the $\mathrm{Y}_{sort}$ across arrays

This algorithm gives each array the same distribution by calculating the mean of each quantile and substituting it as the data value in the original data. The advantage of quantile normalization is it can quickly normalize within a set of chips the same time, without choosing either a baseline array to which all other arrays are normalized or working in a pair-wise manner.

It could be dealt reliably to non-linearity. The main drawback of this method is the strong assumption that the distributions of array intensities are identical, even if individual probes differ in their positions in the distribution. This is true for low abundance genes, and to a fairly good approximation for genes of moderate abundance, but certainly not true for the few high-abundance genes, whose typical levels vary noticeably from sample to sample. Also since it forces the values of quantile to be equal, this could be potentially problematic in the tails where it is possible that a probe could have same values across all the arrays.

Here is an illustration of quantile normalization.

Given a matrix: $\begin{pmatrix} 1 & 5 & 3 & 5 \\ 2 & 1 & 6 & 7 \\ 3 & 2 & 2 & 6 \\ 4 & 6 & 1 & 8 \end{pmatrix}$

Step 1: Sort each column of the original matrix X generate a sorted matrix Y

$$\begin{pmatrix} 1 & 5 & 3 & 5 \\ 2 & 1 & 6 & 7 \\ 3 & 2 & 2 & 6 \\ 4 & 6 & 1 & 8 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{pmatrix}$$

Step 2: Take averages across rows and generate a vector $Y_{sort}$

$$\begin{pmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{pmatrix} \rightarrow \begin{pmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{pmatrix}$$

Step 3: Set average as values for all elements in a row

$$\begin{pmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{pmatrix} \rightarrow \begin{pmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{pmatrix}$$

Step 4: Unsort columns of matrix to the original order

$$\begin{pmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{pmatrix} \rightarrow \begin{pmatrix} 2 & 4.5 & 4.5 & 2 \\ 3 & 2 & 6 & 4.5 \\ 4.5 & 3 & 3 & 3 \\ 6 & 6 & 2 & 6 \end{pmatrix}$$

### 2.2.3 Lowess Smoothing Curves

Lowess stands for "LOcally WEighted polynomial regreSSion", which is a process considered local because each fitted value is determined by neighboring data points defined within a specified span. At each point in the data set, a polynomial is fitted in a subset, with explanatory independent variable values near this point whose outcome is being estimated. The extent of the neighborhood is defined by the "span" parameter which defines the proportion of points out of the total sample size to consider in the neighborhood.

The ideas of local smoothing date back to 1979 when Cleveland[28] originally proposed the method which focuses in a univariate regression setting with one independent variable. Later in 1988[29], Cleveland expanded this idea to multivariate observations.

Lowess is also a weighed process because the regression weight function is defined for the data points contained within the span. The polynomial is fitted by weighted least squares, giving more weight to points near whose response is being estimated, and less weights to points further away. The values of the function for this point are obtained by evaluating the local polynomial using the explanatory variable value. For each data points(x,y) and a given span, the regression weight is calculated by the tricube function:

$$w_i = (1 - |\frac{x - x_i}{d(x)}|^3)^3$$

Where $x$ is the predictor value, $x_i$ is the nearest neighbors of $x$ as defined by the span, and $d(x)$ is the distance from $x$ to the most distant predictor value within the span. Be reminded that in lowess curve procedure, the data points to be smoothed have the largest weight and the most influence on the fit. Data points outside the span have 0 weight so no influence on the fit. Then the coefficients in regression are estimated by minimizing the following function using the least square method:

$$\sum_{k=1}^{n} w_k(x_i)(y_k - \beta_0 - \beta_1 x_k)^2$$

Here $y_k$ is the $k^{th}$ value of the dependent variable, $x_k$ is the $k^{th}$ value of the independent variable. $w_k(x_i)$ is the weight of each point with respect to the center $x_i$, $\beta_0$ and $\beta_1$ are the intercept and regression coefficient, and $n$ is the sample size.

Dudoit and Speed et. al [17] performed a within print-tip group intensity dependent normalization using the scatterplot smoother implemented in the lowess() function from the Splus software (Venables and Ripley, 1996 [30]):

$$\log_2 R/G \rightarrow \log_2 R/G - \mathrm{c}_j(A) = \log_2 k_j(A) R/G$$

Where $\mathrm{c}_j(A)$ is the lowess() fit to the M.A plot. They used f = 20 % to 40% for the parameter specifying the data fraction using for smoothing at each point. Since a very small proportion of genes are expected to vary in expression between red and green labeled mRNA samples, the normalization is uses upon all the genes. In other circumstances, the housekeeping genes could be selected for normalization purpose (Yang et al [26, 31]). In R, lowess curve is implemented by “loess” package.

# Chapter 3: Normalization for a Lung Cancer Marker Dataset

## 3.1 Study illustrations

Lung cancer is the leading cause of cancer-related deaths in the United States[32]. Cigarette smoking is the most important risk factor for lung cancer and accounts for 85-90% of lung cancer cases [33-34]. Although smoking prevalence has decreased as a result of emphasis on prevention, lung cancer continues to be a major problem partly due to persistent risk among former smokers[35]. In the United States, lung cancer is diagnosed more commonly in former than current smokers. During many studies of lung cancer, some findings support the concept that interactions between tumor cells and other components of tumor microenvironment may produce a unique molecular protein profile. If possible, this protein profile could be utilized as a diagnostic test.

CT scan screening of nodules in the lung may not resolve the issue of whether nodules are malignant or benign. If nodules are large with many spots, they are sure to be malignant. If nodules are small, they are probably benign. But the interpretation of intermediate sized nodules is more complicated and patients may have to undergo surgery to have nodules biopsied and/or removed.

The plasma data come from a study done by the American College of Surgeons Oncology Group (ACOSOG). This study utilizes the biospecimen resources from Z4031 trial and has been assigned study number Z4093. Trial Z4031 includes a biospecimen repository of serum specimens from patients with suspicious lung lesions that underwent surgery for nodule resection. A subset of these patients had benign pulmonary nodules and thus serves as an ideal control group for the proposed research study (993 total samples; 799 serum specimens from patients

with lung cancer and 194 serum specimens from patients without cancer). Peripheral blood was obtained before and after lung nodule resection at 90 days post-operatively. The study receives and tests these specimens in a blinded manner with the cooperative group holding the outcome data.

We requested and obtained plasma samples both before and post operation from 50 patients with diagnosed malignant nodules and single plasma samples from 50 patients with benign nodules; 150 plasma samples in all. To assess whether a marker panel could discriminate between patients with malignant and benign nodules, a panel of antibodies against 48 potential markers composed of proteins hypothesized to contribute to lung cancer progression was assembled. All lung cancer and control samples were collected and processed utilizing a standardized collection and storage protocol previously used by the NIH/NHLBI sponsored Lung Health Study trial (LHS).

In order to prevent experimental artifacts that could arise from analyzing cancer samples and non- cancer samples separately, the study analyzes cancer and control serums together on each assay plate, with samples randomized across plates in the assay system. In addition, 8 standards and 1 control plasma are included on each plate to provide a means of normalizing results between plates run on separate days.

However, from the earlier analysis, there is no statistically significant difference of intensities between normal and cancer samples, which seem to suggest that markers are not useful to detect the lung cancer early. But other than markers themselves, this might due to other reasons such as experimental variability. We would like to explore the variability across plates for each marker before we conclude for markers. This motivates us to explore this study further.

## 3.2 Quantification

A bead based multiplex immunoassay utilizing a Luminex assay system was used to evaluate the effects of these markers.

Data on 48 markers were to be measured for each of the 150 plasma samples. Two standard assay plates were used to provide data for each sample, a 21-plex marker plate and a 27-plex marker plate. Each plate has space for 96 samples. Each of the plasma samples was measured in triplicate, for a total of 450 "samples" divided across eight 21-plex and eight 27-plex plates. Samples from patients with and without malignant nodules were randomized to each plate, each was measured in triplicates on that plate and on the average 19 samples per plate. Each plate also contained one blank (duplicates), one "normal" (duplicates or triplicates), and eight singletons standards which served as a basis for estimating the standard curve, a four-parameter logistic. The samples arranged the same on both the 21-plex and 27 plex plates.

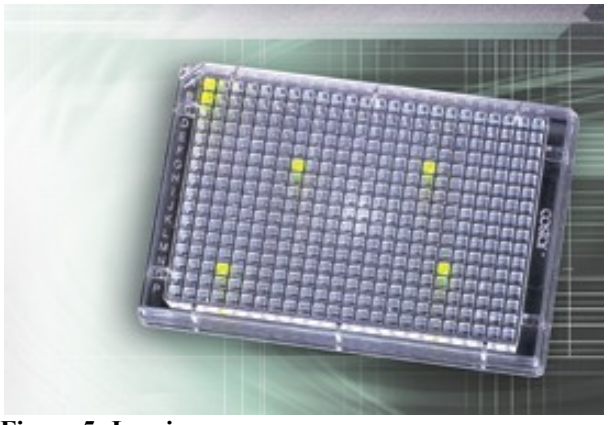

**Figure 5: Luminex assay**

The Luminex's xMAP® system is built on proven, existing technology of flow cytometry. Featuring a flexible, open-architecture design, xMAP technology can be configured to perform a wide variety of bioassays quickly, cost-effectively and accurately, which require only a miniscule amount of sample (<100 l). First, Luminex® color-codes tiny beads, called

microspheres, into 100 distinct sets. Each bead set can be coated with a reagent specific to a particular bioassay, allowing the capture and detection of specific analytes from a sample. Within the Luminex analyzer, lasers excite the internal dyes that identify each microsphere particle by color, and also any reporter dye captured during the assay. Many readings are made on each bead set, further validating the results. In this way, xMAP technology allows multiplexing of up to 100 unique assays within a single sample, both rapidly and precisely. In order to prevent experimental artifacts from corrupting the data we randomize all samples (cancer and control groups) across the assay plates. In addition all samples were run in triplicate and these replicates were also randomized across the assay plates.

The data structures within the 8 plates are as follows: the average of replicates is shown in the following.

| | Marker1 | Marker2 | …… | Marker n |
|---|---|---|---|---|
| Blank | ● | ● | …… | ● |
| Standard1 | ● | ● | …… | ● |
| …… | …… | …… | …… | …… |
| Standard8 | ● | ● | …… | ● |
| Control | ● | ● | …… | ● |
| Sample1 | ● | ● | …… | ● |
| …… | …… | …… | …… | …… |
| Sample18 | ● | ● | …… | ● |

**Table 2: Plate Layout of 21-Plex**

Three measurements are made for each sample, each marker and each plate--fluorescent intensity (FI), background in FI and observed concentration (OC). Each measurement is made at two scanning intensities, either high or standard. Therefore there are six different measurements to consider: high vs. low FI, high vs. low background in FI, high vs. low OC.

The first question we want to tackle is: which one of the 6 measurements is in tightest estimation of the true markers for the control plasma/standard plasma matrix. Since control or standards plasma are from the same sample, in principle we expect to read similar values from each plate. These values are expressed by mean plus some experimental variability across plates. For different markers, there might be different schemes to measure them best. However we would like to identify first if there is one measurement that gives lowest coefficient of variation and most precision across plates consistently. Then this is the measurement we would like to quantify before proceeding normalization. Once the best method of measurement is chosen, we can proceed to determine which of the 8 standards or the control plasma is most suitable.

We started with the first 21 markers on the control plasma. We extracted control plasma from each plate and repeat this across 8 plates. Thus it makes a matrix of 8*21 where row is plate and column is marker. Since there are six different measurements, this composes six separate control plasma matrices. The coefficient of variation across plates is calculated from each control plasma matrix, and the comparison on the six measurements has been made. Repeat the same procedure for the remaining 27 markers.

One criterion to use in choosing a measurement method to use is to select the one for which plate to plate variability is the least. Here we measure plate to plate variability for an outcome measure by computing the coefficient of variation across the eight plates for the “normal control”

and for the "standard" samples for each marker. The coefficient of variation (c.v.) is the ratio of the standard deviation across the eight plates to the mean across the eight plates. We choose the coefficient of variation rather than the raw standard deviation because of the large scale variation in intensity across markers.

Tables 2 and 3 show the coefficient of variation for the six measurement methods for the eight 21 marker plates and for the eight 27 marker plates for the "normal control". For the 21 markers, the c.v. is the smallest for standard FI for 12 of the 21 markers; the average c.v. across markers for standard F.I. is 0.21. For the remaining 27 markers, the lowest c.v. is usually provided by using the standard or high FI; the means across markers are 0.16 and 0.19 respectively (Table 3).

| Marker: Coefficient of variation | High FI | High FI – background | High observed conc | Standard FI | Standard FI - background | Standard observed conc |
|---|---|---|---|---|---|---|
| Marker:1 | **0.13** | 0.25 | 0.64 | 0.15 | 0.14 | 0.39 |
| Marker:2 | 0.48 | 1.23 | 0.17 | **0.10** | 0.39 | 0.14 |
| Marker:3 | 0.12 | 0.63 | 1.98 | **0.09** | 0.82 | 0.17 |
| Marker:4 | 0.26 | 0.73 | 0.27 | 0.17 | 0.26 | **0.15** |
| Marker:5 | 0.24 | 0.76 | 0.26 | **0.16** | 0.21 | 0.20 |
| Marker:6 | 0.22 | 1.03 | 0.32 | **0.17** | 0.45 | 0.28 |
| Marker:7 | 0.23 | 0.43 | 0.14 | 0.15 | 0.16 | **0.12** |
| Marker:8 | 0.26 | 0.24 | 0.19 | 0.20 | 0.21 | **0.12** |
| Marker:9 | 0.18 | 0.21 | 0.65 | 0.13 | **0.12** | 0.65 |
| Marker:10 | 0.24 | 0.60 | 1.91 | **0.21** | 0.30 | 1.24 |
| Marker:11 | **0.17** | 0.33 | 0.76 | 0.18 | 0.31 | 0.68 |
| Marker:12 | 0.20 | 0.25 | 0.17 | 0.16 | 0.18 | **0.14** |
| Marker:13 | 0.47 | 0.36 | 0.34 | **0.16** | 0.17 | 0.19 |
| Marker:14 | 0.25 | 0.45 | 0.39 | **0.21** | 0.26 | 0.34 |
| Marker:15 | 0.18 | 0.40 | 0.11 | 0.18 | 0.21 | **0.09** |
| Marker:16 | 0.24 | 0.50 | 0.31 | **0.17** | 0.17 | 0.30 |
| Marker:17 | 0.22 | 0.46 | 0.41 | **0.13** | 0.14 | 0.15 |
| Marker:18 | 0.18 | 1.24 | 0.52 | **0.13** | 0.67 | 0.57 |
| Marker:19 | 0.16 | 0.29 | 0.16 | **0.15** | 0.19 | 0.34 |
| Marker:20 | 0.23 | 0.22 | **0.11** | 0.14 | 0.17 | 0.68 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Marker:21 | 0.24 | 0.52 | 0.48 | **0.18** | 0.24 | 0.59 |

**Table 3: Coefficient of variation across 8 plates for 6 measurements on "normal control" (21-plex)**

| Marker: Coefficient of variation | High FI | High FI minus background | High observed concentration | Standard FI | Standard FI minus background | Standard observed concentration |
|---|---|---|---|---|---|---|
| Marker:22 | 0.09 | **0.08** | 0.08 | 0.09 | 0.09 | 0.09 |
| Marker:23 | 0.24 | 0.36 | 0.60 | **0.22** | 0.35 | 0.49 |
| Marker:24 | 0.26 | 0.33 | 0.31 | **0.24** | 0.29 | 0.36 |
| Marker:25 | 0.24 | 0.63 | N/A | **0.22** | 1.05 | N/A |
| Marker:26 | **0.18** | 0.34 | 0.16 | 0.27 | 0.36 | 0.28 |
| Marker:27 | 0.31 | 0.40 | 0.27 | 0.22 | 0.28 | **0.13** |
| Marker:28 | **0.22** | 0.37 | 0.36 | 0.24 | 0.37 | 0.32 |
| Marker:29 | **0.22** | 0.27 | 0.32 | 0.24 | 0.28 | 0.39 |
| Marker:30 | **0.16** | 0.23 | 0.65 | 0.26 | 0.36 | 0.81 |
| Marker:31 | 0.23 | 0.83 | 0.54 | **0.21** | 0.84 | 0.39 |
| Marker:32 | 0.22 | 0.82 | 0.88 | **0.20** | 1.86 | 0.72 |
| Marker:33 | **0.13** | 0.34 | 0.60 | 0.26 | 0.50 | 0.41 |
| Marker:34 | 0.21 | 0.30 | 0.18 | 0.22 | 0.28 | **0.20** |
| Marker:35 | 0.15 | 0.43 | N/A | **0.13** | 0.63 | N/A |
| Marker:36 | 0.14 | 2.73 | 0.58 | **0.14** | 2.72 | 0.74 |
| Marker:37 | 0.10 | 0.10 | 0.07 | 0.12 | 0.11 | **0.06** |
| Marker:38 | **0.14** | 0.81 | N/A | 0.16 | 0.98 | 0.63 |
| Marker:39 | **0.25** | 0.43 | 0.39 | 0.31 | 0.48 | 0.45 |
| Marker:40 | **0.19** | 1.82 | 0.72 | 0.20 | 2.40 | 0.51 |
| Marker:41 | 0.26 | 0.47 | 0.32 | **0.18** | 0.23 | 0.21 |
| Marker:42 | 0.06 | **0.06** | 0.07 | 0.09 | 0.09 | 0.07 |
| Marker:43 | 0.14 | 0.30 | 0.11 | 0.19 | 0.27 | **0.11** |
| Marker:44 | **0.13** | 1.02 | 0.32 | 0.21 | 1.27 | 0.52 |
| Marker:45 | 0.11 | **0.10** | 0.11 | 0.13 | 0.12 | 0.10 |
| Marker:46 | **0.05** | 0.05 | 0.07 | 0.06 | 0.06 | 0.07 |
| Marker:47 | 0.28 | 1.92 | 0.52 | **0.21** | 3.32 | 0.91 |
| Marker:48 | 0.14 | 0.88 | N/A | **0.20** | 0.92 | 0.93 |

**Table 4: Coefficient of variation across 8 plates for 6 measurements on "normal control" (27-plex)**

| Average c.v. | High FI | High FI minus background | High observed concentration | Standard FI | Standard FI minus background | Standard observed concentration |
|---|---|---|---|---|---|---|
| 21-marker | 0.22(0.09) | 0.52 (0.32) | 0.70(1.03) | 0.15(0.03) | 0.27(0.18) | 0.42(0.49) |
| 27-marker | 0.16(0.06) | 0.61(0.63) | 0.36(0.24) | 0.19(0.06) | 0.76(0.86) | 0.40(0.27) |

**Table 5: Average coefficient of variation across plates on "normal control"**

In addition, background in FI is calculated by subtracting background from FI. According to the definition of the coefficient of variation, background in FI across plates would share the same standard deviation with FI, but with smaller mean. This results in the larger coefficient of variation for the background in FI. Therefore, we compared the standard deviation rather than coefficient of variation between FI and background in FI. Table 4 shows that high or standard FI is more consistent across plates than background, as it has smaller standard deviation.

| Standard Deviation | High FI | High FI minus background | Standard FI | Standard FI minus background |
|---|---|---|---|---|
| 21-marker | 160.38 | 228.65 | 30.78 | 31.44 |
| 27-marker | 85.94 | 98.13 | 24.70 | 26.24 |

**Table 6: Standard deviation across plates for "normal control"**

Furthermore, we notice that there are many "<< OOR" or ">> OOR" in observed concentration. These concentrations are read out of range of detection, which is determined based on the standard curve. The simplest method for determining concentrations from a standard curve is to construct a plot of fluorescent intensity vs. concentration by linear regression[36] . In this project, a four parameter logistic (4PL) has been used to construct the standard curve. The 4PL equation contains 4 parameters in the curve, as illustrated in Figure 7.

$$y = d + \frac{a - d}{[1 + (\frac{x}{c})^b]}$$

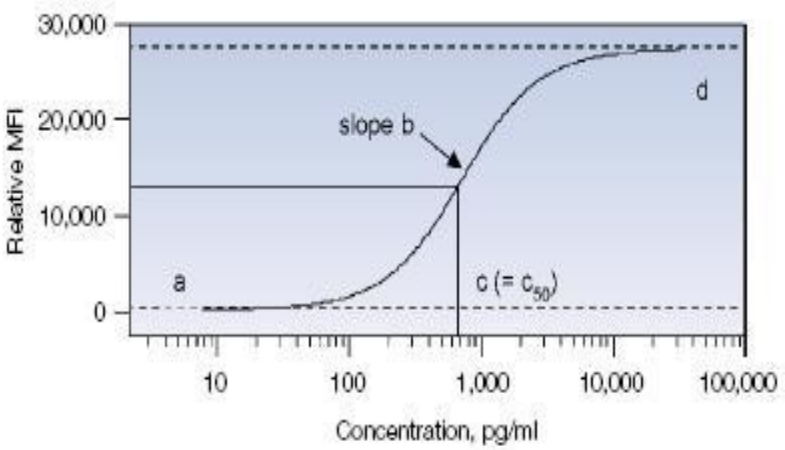


**Figure 6: Standard curve fitted with 4-PL regression**

Where a is the estimated response at zero concentration, b is the slope factor, c is the mid-range concentration ($C_{50}$), and d is the estimated response at infinite concentration. In principle, once the 4PL equation is created from a set of standards, the 4 parameters could be determined. Then the equation could be used to calculate unknown concentrations (x) from the fluorescent intensity (y). However, in this dataset, 8 standards in each plate are used to draw the standard curve. Among these standards, two of them are saturated, and two of them are responded at very low concentration. Literally only 4 standards in the range of samples are appropriate to estimate the standard curve. However the 4PL standard curve is composed of four parameters. Thus only the remaining four standards can potentially be used to estimate the four parameters in the standard curve. Statistically speaking, this results in over-fitting. This suggests that the middle four standards, 3-6, will generally be most informative. The observed concentrations derived from the standard curve are not reliable.

On the other hand, ideally fluorescent intensity falls into the middle range of the standard curve (sensitive region). If fluorescent intensity is higher than the sensitive region, we call it is saturated. If the fluorescent intensity is too low, the concentration extracted from the curve is not

accurate. However samples out of the range of the standard curve are not necessarily out of limits of detection, but probably are out of limits of quantification by the standard curve when machine cannot read these numbers. We counted the number of OOR in samples for each marker. On the average, 21-plex has 40 out of 150 samples that are out of range, and 27-plex has 58. Practically there is disadvantage of using observed concentration.

Once we decide to select standard fluorescent intensity as the best measurement, we proceeded to determine whether standards/control plasma provide consistent behavior across plates. We ranked each standard /control plasma across plates for each marker and averaged the rankings across markers. We then calculated the variance of the ranking across 8 plates for standards and control plasma respectively. The rank represents the consistency of the order preservation. The maximum variance of the ranking across 8 plates would be 6. Normal plasma has the highest variance which is 4.19. This suggests that normal plasma has the best separation of ranking across plates. S1 and S6 performed comparably well to the performance of control plasma. However S1 is mostly saturated and out of the sample range. Other standards have smaller variance suggesting they are provide less consistent order preservation across plates (Table 4). Overall standard 6 and control plasma provide consistent separation and perform better than other standards.

| | Variance of Ranking |
|---|---|
| Normal Plasma | 4.19 |
| Standard 1 | 4.02 |
| Standard 2 | 2.71 |
| Standard 3 | 3.13 |

| | |
|---|---|
| Standard 4 | 2.75 |
| Standard 5 | 1.97 |
| Standard 6 | 4.21 |
| Standard 7 | 3.17 |
| Standard 8 | 3.71 |

**Table 7: Variance of ranking across 8 plates for control plasma and standards**

In brief, since standard fluorescent intensity is a consistent measurement providing small coefficient of variation, it is chosen as the measurement we will utilize for further analysis. In addition we will evaluate normalization methods by examining their effects on the "normal control" and Standard 6.

## 3.3 Normalization

In the following we evaluate the effects of four different normalization methods: scale normalization (comparison to "baseline" plate using median, mean, or third quartile), quantile normalization, lowess smoothing curve normalization, and lowess curve extrapolation.

We first introduce the median sample matrix to serve as the basis for each of the normalization methods and the assessment matrices using Sample A and/or Sample B.

Median sample matrix is used to calculate normalization factor or fit the lowess curve. Sample A and/or Sample B matrices are used to evaluate the coefficient of variation across plates.

The median sample matrix is constructed by taking the median value for each marker across the clinical samples on each plate (blanks, standards, and normal control are excluded). Thus there is a pair of median sample matrices--an 8 x 21 (plate x marker) matrix and an 8 x 27 (plate x marker) matrix.

Two pairs of assessment matrices are also constructed. One pair contains the "normal control" (Sample A values) for the 21-plex and the 27-plex plates, one pair contains the values for Sample B, uses a mixture of "normal control" and Standard 6 values. For Sample B for each plate and each marker we choose whichever of the "normal control" or the Standard 6 values is closest to the median across clinical samples.

Sample A matrix is organized as follows:

1) Within each plate, choose control plasma intensity that is labeled as "C1"
2) Repeat this for all the 8 plates, then organize them as a control plasma matrix
3) The dimension for the control plasma is 8 × 21 or (8 × 27), where rows represent plates from 1 to 8, and columns represent each marker.

Similarly, sample B matrix is constructed by:

1) Within each plate, choose median intensities across samples
2) Since control plasma and standard 6 give consistent separation of plates, we choose intensity from either control plasma or standard 6 that is closest to median samples
3) Following this procedure for all the 8 plates, the standard plasma matrix is organized.
4) The dimension for the standard plasma is 8 × 21 or (8 × 27), where rows are plates and columns are markers.

If plate to plate variation is reduced by a normalization method the normalized assessment matrices should show low plate to plate variability for each marker.

### 3.3.1Median Normalization

These scale normalization methods are based on adjusting values for each array to values for a selected baseline/comparison array. The disadvantage of scale normalization is that it depends on the arbitrary choice of a baseline array.

The first step in these scale normalization methods is to choose a baseline plate. We usually choose a plate whose marker values tended to fall close to the median across all plates. This plate was labeled plate 1.

Then for median normalization, each expression value, marker *j* on plate *i* for sample *k*, is normalized by multiplying it by the factor (median expression for marker *j* on baseline plate 1) / (median expression for maker *j* on plate *i*). Mean and third quartile normalizations are done in the same way except that the mean or the third quartile value is used instead of the median in the normalization factor.

To evaluate the normalization approaches, the normalizations were applied to Sample A and Sample B values for each marker and plate. Then the coefficient of variation was computed across plates for each marker and the ratio of the coefficient of variation for the unnormalized compared to the coefficient for the normalized values computed for each marker. See results in Table 5 below.

Table 5 and 6 show that when we apply median normalization factors back to the Sample A matrix, most c.v. ratios are above 1, and a few ratios are smaller but very close to 1. The mean ratio of c.v. between unnormalized and normalized for 21-marker is 1.02, and for 27-marker is 1.23. This demonstrates the median normalization has made some reduction in variability. Although the ratios are not too far away from 1, median normalization reduced the variation across plates at a certain degree. Other than median normalization, mean and Q3 normalization showed similar results.

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|
| 1 | 1.01 | 12 | 1.04 |
| 2 | 0.99 | 13 | 0.99 |
| 3 | 0.99 | 14 | 1.02 |
| 4 | 1.05 | 15 | 1.00 |

| | | | |
|---|---|---|---|
| 5 | 0.99 | 16 | 1.00 |
| 6 | 1.01 | 17 | 1.04 |
| 7 | 1.06 | 18 | 1.03 |
| 8 | 1.01 | 19 | 1.06 |
| 9 | 1.05 | 20 | 1.02 |
| 10 | 1.04 | 21 | 1.00 |
| 11 | 1.01 | | |

**Table 8: Median normalization to Sample A: ratio of coefficient of variation (CV) between unnormalized and normalized for the first 21 markers**

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|---|---|
| 1 | 2.36 | 10 | 1.14 | 19 | 1.07 |
| 2 | 1.11 | 11 | 1.20 | 20 | 1.19 |
| 3 | 1.13 | 12 | 1.19 | 21 | 0.95 |
| 4 | 1.16 | 13 | 1.13 | 22 | 1.25 |
| 5 | 1.07 | 14 | 1.26 | 23 | 1.21 |
| 6 | 1.12 | 15 | 1.47 | 24 | 1.36 |
| 7 | 1.14 | 16 | 1.63 | 25 | 1.47 |
| 8 | 1.03 | 17 | 1.23 | 26 | 1.09 |
| 9 | 1.04 | 18 | 1.02 | 27 | 1.22 |

**Table 9: Median normalization to Sample A: ratio of coefficient of variation (CV) between unnormalized and normalized for the remaining 27 markers**

For Sample B, the mean ratio of c.v. between unnormalized and normalized on standard plasm matrix for 21-marker is 1.30, and for 27-marker is 1.05. The median normalization reduces more variability in Sample B than Sample A matrix.

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|
| 1 | 1.52 | 12 | 1.03 |
| 2 | 1.10 | 13 | 1.19 |
| 3 | 1.18 | 14 | 1.28 |
| 4 | 1.13 | 15 | 1.43 |
| 5 | 0.96 | 16 | 1.00 |
| 6 | 1.48 | 17 | 1.91 |
| 7 | 1.11 | 18 | 2.59 |
| 8 | 1.06 | 19 | 1.12 |
| 9 | 1.97 | 20 | 0.97 |
| 10 | 0.85 | 21 | 1.13 |
| 11 | 1.35 | | |

**Table 10: Median normalization to Sample B matrix: ratio of coefficient of variation (CV) between unnormalized and normalized for the first 21 markers**

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|---|---|
| 1 | 0.95 | 10 | 1.08 | 19 | 1.07 |
| 2 | 0.94 | 11 | 1.01 | 20 | 1.19 |
| 3 | 1.12 | 12 | 1.05 | 21 | 0.95 |
| 4 | 1.22 | 13 | 1.23 | 22 | 1.30 |
| 5 | 1.01 | 14 | 0.86 | 23 | 1.21 |
| 6 | 0.94 | 15 | 1.16 | 24 | 0.90 |
| 7 | 0.93 | 16 | 0.99 | 25 | 0.68 |
| 8 | 1.03 | 17 | 1.01 | 26 | 1.06 |
| 9 | 1.22 | 18 | 1.02 | 27 | 1.06 |

**Table 11: Median normalization to Sample B matrix: ratio of coefficient of variation (CV) between unnormalized and normalized for the remaining 27 markers**

| Median Normalization | Mean Ratios of c.v. |
|---|---|
| 21-marker from Sample A matrix | 1.02 |
| 27-marker from Sample A matrix | 1.23 |
| 21-marker from Sample B matrix | 1.30 |
| 27-marker from Sample B matrix | 1.05 |

**Table 12: Median normalization: mean ratios of c.v. among Sample A and B matrices**

### 3.3.2 Quantile Normalization

After the median normalization, we made an attempt to normalize the Sample A and B matrix with quantile normalization, by first applying the quantile normalization on the median sample matrix.

Upon the normalization, intensity in median sample matrix is matched up across all 8 plates so that the smallest median value on each array is identical, the second smallest is identical, and so forth. Figure 8 shows an example of such intensities on plates 5 and 6, before and after the quantile normalization. Before normalization, intensities in plate 5 and 6 are scattered around diagonal line but not exactly corresponding to each other. After the normalization, intensities are aligned exactly on the diagonal line, showing plate 5 is matched identically with plate 6. Figure 9 is the box plot of eight plates before and after quantile normalization. There was substantial variability across plates before quantile normalization, but this variability was removed by the quantile normalization.

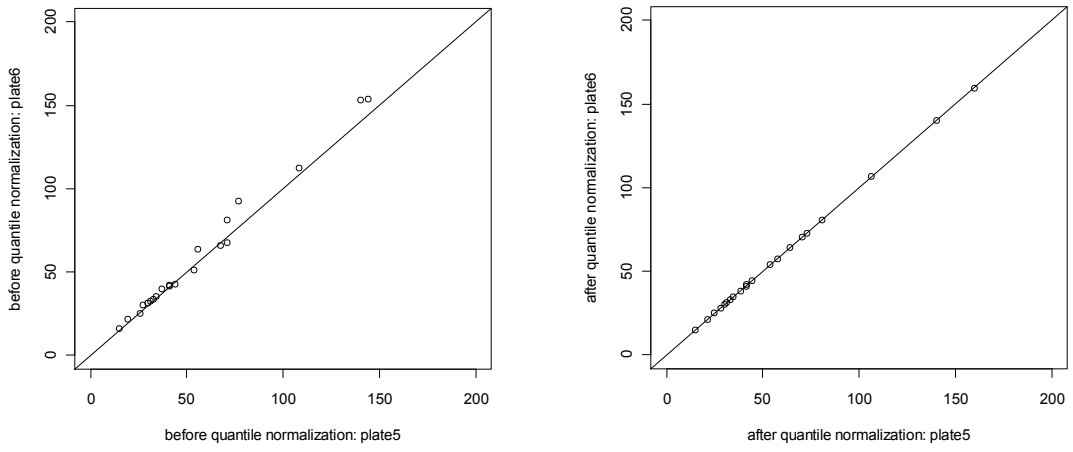


**Figure 7: before and after quantile normalization: mean intensity in plate 5 vs. plate 6**

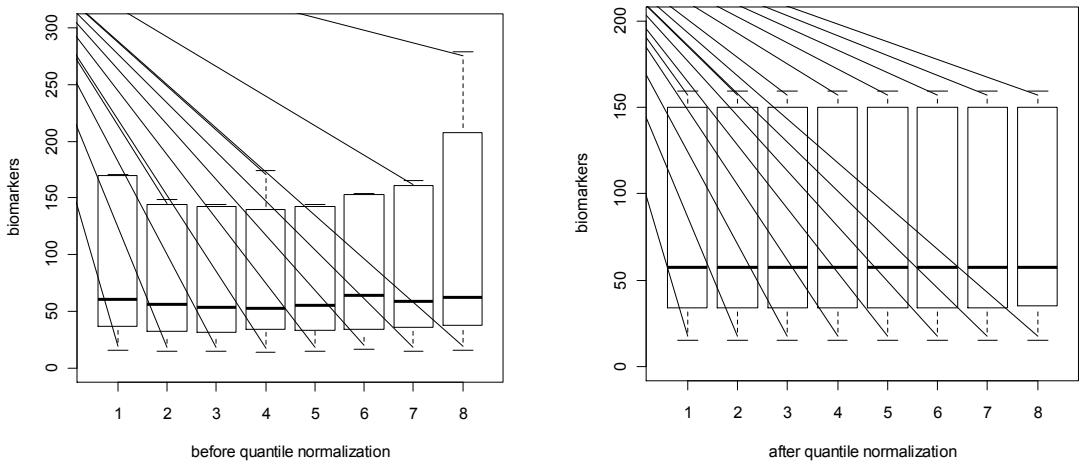


**Figure 8: box plots for eight plates before and after quantile normalization**

After the median intensity in each plate is normalized, the deviation is obtained by subtracting the normalized intensity from the median intensity in each plate. Then we apply the deviation back to Sample A or Sample B, by subtracting deviation from intensities in assessment matrix to obtain the normalized values. The ratio of coefficient of variation between unnormalized and normalized Sample A/B is compared.

For 21 marker in Sample A matrix, the ratios of c.v. from markers 2, 11, 12, 13, 16, and 20 are smaller than 1. Particularly markers 2, 12, 13 and 20's ratios are smaller than 0.5. For 27-marker, 20 out of 27 (74%) markers whose ratios of c.v. are smaller than 0.5. For Sample B, it shows more number of markers with ratio of c.v. greater than 1. However in 21-marker, marker 13's ratio is equal to 0.61. In 27-marker system, 15 out of 27 (55.56%) markers' ratios are smaller than 0.5.Quantile normalization is not consistent in reducing the variation across plates.

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|
| 1 | 1.07 | 12 | 0.52 |
| 2 | 0.30 | 13 | 0.35 |
| 3 | 1.46 | 14 | 1.20 |
| 4 | 1.19 | 15 | 1.63 |

| | | | |
|---|---|---|---|
| 5 | 1.22 | 16 | 0.78 |
| 6 | 1.20 | 17 | 1.02 |
| 7 | 1.26 | 18 | 1.76 |
| 8 | 1.28 | 19 | 1.58 |
| 9 | 1.34 | 20 | 0.21 |
| 10 | 1.31 | 21 | 1.17 |
| 11 | 0.85 | | |

**Table 13: Quantile normalization for the first 21 markers: ratio of coefficient of variation (CV) between unnormalized and normalized Sample A matrix**

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|---|---|
| 1 | 0.10 | 10 | 0.39 | 19 | 0.18 |
| 2 | 0.18 | 11 | 0.74 | 20 | 0.16 |
| 3 | 2.40 | 12 | 0.42 | 21 | 0.20 |
| 4 | 0.17 | 13 | 0.18 | 22 | 0.27 |
| 5 | 1.70 | 14 | 0.13 | 23 | 0.63 |
| 6 | 0.24 | 15 | 0.37 | 24 | 0.48 |
| 7 | 0.20 | 16 | 0.30 | 25 | 0.64 |
| 8 | 0.33 | 17 | 0.55 | 26 | 0.92 |
| 9 | 0.35 | 18 | 1.20 | 27 | 0.30 |

**Table 14: Quantile normalization for the remaining 27 markers: ratio of coefficient of variation (CV) between unnormalized and normalized Sample A matrix**

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|
| 1 | 1.07 | 12 | 1.09 |
| 2 | 1.46 | 13 | 0.61 |
| 3 | 1.46 | 14 | 1.20 |
| 4 | 0.99 | 15 | 1.62 |
| 5 | 1.13 | 16 | 1.14 |
| 6 | 1.57 | 17 | 1.02 |
| 7 | 1.15 | 18 | 1.76 |
| 8 | 1.19 | 19 | 1.18 |
| 9 | 1.47 | 20 | 0.92 |
| 10 | 0.88 | 21 | 1.06 |
| 11 | 0.88 | | |

**Table 15: Quantile normalization for the first 21 markers: ratio of coefficient of variation (CV) between unnormalized and normalized Sample B matrix**

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|---|---|
| 1 | 1.09 | 10 | 0.31 | 19 | 0.08 |
| 2 | 0.19 | 11 | 0.74 | 20 | 0.17 |

| | | | | | |
|---|---|---|---|---|---|
| 3 | 0.34 | 12 | 1.13 | 21 | 0.54 |
| 4 | 0.19 | 13 | 0.18 | 22 | 0.21 |
| 5 | 0.53 | 14 | 0.13 | 23 | 0.63 |
| 6 | 0.20 | 15 | 0.30 | 24 | 0.73 |
| 7 | 0.16 | 16 | 1.25 | 25 | 0.87 |
| 8 | 0.28 | 17 | 0.55 | 26 | 0.87 |
| 9 | 0.35 | 18 | 0.49 | 27 | 0.88 |

**Table 16: Quantile normalization for the remaining 27 markers: ratio of coefficient of variation (CV) between unnormalized and normalized Sample B matrix**

### 3.3.3 Lowess Curve Normalization

For each marker in the median sample matrix, median across 8 plates is selected so that a median mock array ($1 \times 27$) is fashioned out of the averages of the plates being normalized. The normalization procedure is carried out as follows:

1) Upon median sample matrix, fit 8 smoothing curves, where each fit is intensity in each plate vs. median mock array

2) The deviation is calculated from fitted value in each plate minus median mock array

3) Apply the deviation back to Sample A/B matrix, by subtracting the deviation from them

4) Calculate coefficient of variation before and after normalization respectively, and calculate the ratio of the coefficient of variation.

When fitting the lowess curve of plate *i* vs. median mock array, we first check the span parameter. We start from plate 1 in 27-marker first. When the span value is as small as 0.10, the lowess curve cannot be fit. When the span is 0.2, it does not provide much smoothing. When span value is as large as 1 the lowess curve deviates from data points. When span is set as 0.4, 0.6, 0.75, curves are fit well and there is no much difference between them. Overall the default value of 0.75 works fairly well in finding the smoothing curve. This is also applied to other plates.

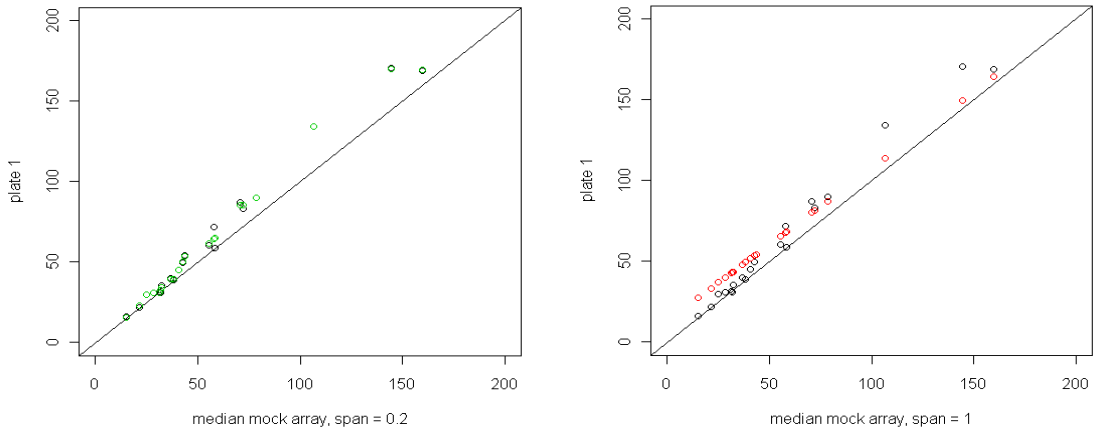


**Figure 9: Lowess curve fitting of plate 1 vs. median mock array (27-plex) when span is 0.2 and 1**

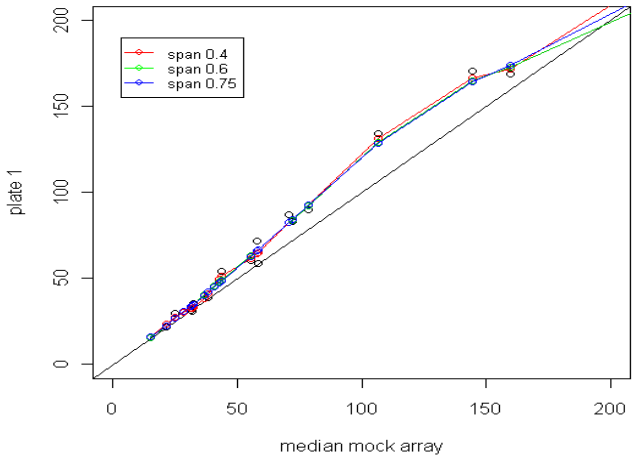


**Figure 10: Lowess curve fitting of plate 1 vs. median mock array (27-plex) when span is 0.4, 0.6 and 0.75**

In this dataset, the sample size for each lowess curve is 21 or 27, since we fit the lowess curve on the 21-marker or 27-marker. The sample size for lowess curve is small. Figure 10 and 11 show the lowess curve fit for each plate (21-marker and 27-marker) in median sample matrix, where y axis is the median sample from each plate and x axis is the median mock array.

For 21-marker, except marker1 whose average intensity is over 2000, others are scaled under 1000. The range of its median mock array for is (14.5, 2144.6). In 27-marker, their intensities

scales are variable across 27 markers. The range of its median mock array is: (15.05, 11745.60), where marker 1's average intensity is 8836, and marker 24's is 11745.

The lowess curve plots have shown most data points are scattered around the left lower corner where they are low in intensity, while a few points are located at right upper side indicating high intensity.

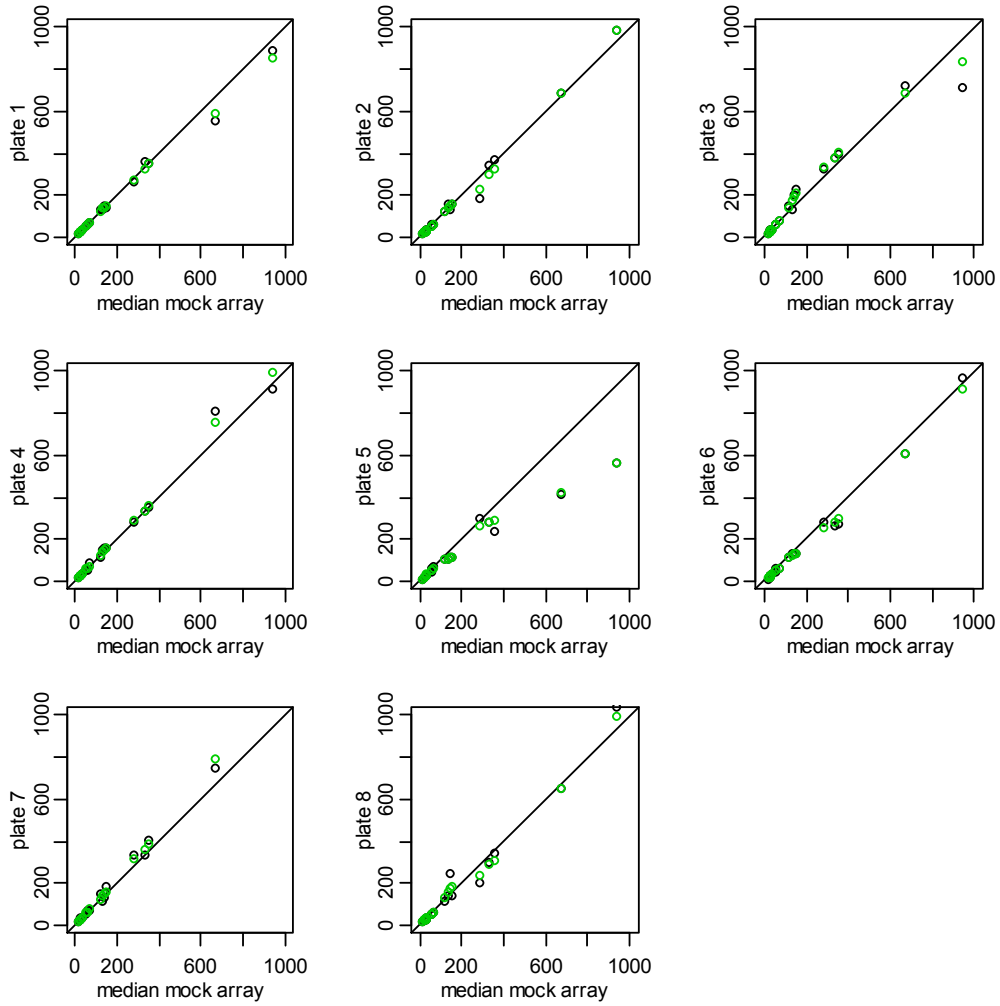


**Figure 11: Lowess curve for 21-marker: plate intensity in median sample matrix vs. median mock array**

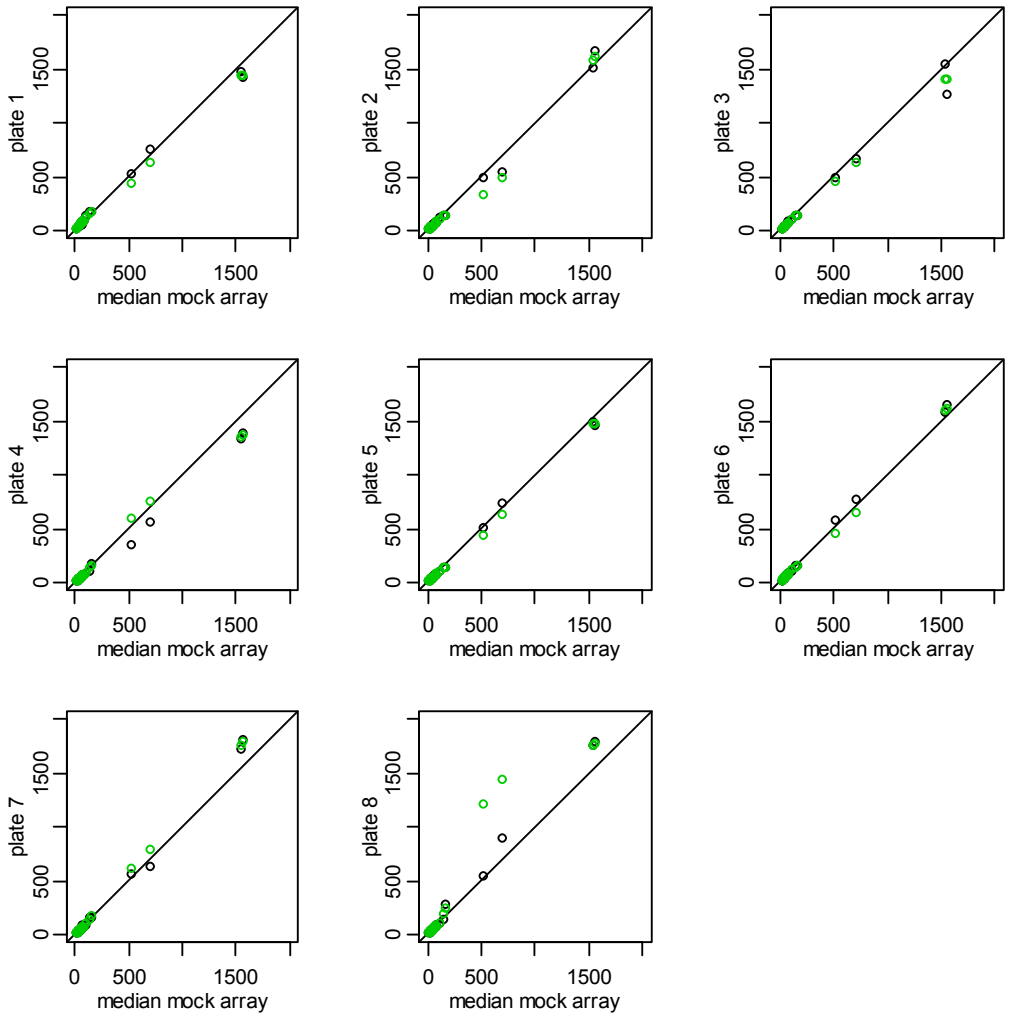


**Figure 12: Lowess curve for 27-marker: plate intensity in median sample matrix vs. median mock array**

Table 14-15 show the ratios of coefficients of variation between unnormalized and normalized in Sample B matrix. For 21-marker, most ratios are greater than 1 except marker 10 and 13. For 27-marker, more markers' ratio are smaller but close to 1. The mean ratios for 21-makrer and 27-marker are 1.20 and 1.05. (Table 16).

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|
| 1 | 1.07 | 12 | 1.05 |
| 2 | 1.01 | 13 | 0.75 |
| 3 | 1.59 | 14 | 1.13 |
| 4 | 1.09 | 15 | 1.63 |

| | | | |
|---|---|---|---|
| 5 | 1.07 | 16 | 1.13 |
| 6 | 1.92 | 17 | 1.20 |
| 7 | 1.17 | 18 | 1.71 |
| 8 | 1.19 | 19 | 1.11 |
| 9 | 1.28 | 20 | 1.05 |
| 10 | 0.78 | 21 | 1.09 |
| 11 | 1.20 | | |

**Table 17: Lowess curve normalization for the first 21 markers: ratio of coefficient of variation (CV) between unnormalized and normalized Sample B matrix**

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|---|---|
| 1 | 0.94 | 10 | 1.06 | 19 | 1.09 |
| 2 | 0.93 | 11 | 1.16 | 20 | 1.00 |
| 3 | 1.35 | 12 | 0.83 | 21 | 0.83 |
| 4 | 1.16 | 13 | 1.14 | 22 | 0.64 |
| 5 | 0.97 | 14 | 1.25 | 23 | 1.23 |
| 6 | 0.95 | 15 | 0.99 | 24 | 0.66 |
| 7 | 0.86 | 16 | 1.95 | 25 | 1.21 |
| 8 | 1.01 | 17 | 1.22 | 26 | 1.05 |
| 9 | 1.36 | 18 | 0.94 | 27 | 0.67 |

**Table 18: Lowess curve normalization for the remaining 27 markers: ratio of coefficient of variation (CV) between unnormalized and normalized Sample B matrix**

| | mean ratios of c.v. |
|---|---|
| 21-marker, standard plasma | 1.20 |
| 27-marker, standard plasma | 1.05 |

**Table 19: mean ratios of c.v. between unnormalized and normalized Sample B matrix**

### 3.3.4 Lowess Curve Extrapolation

Next we developed an extrapolation method utilizing lowess smoothing curve. Again we first fit 8 lowess smoothing curves by median intensity in each plate vs. median mock array. Then by applying the lowess curve back to the control/standard plasma matrix, we back-predict (extrapolate) the lowess curve to obtain the normalized intensity. We then compared the

coefficient of variation between unnormalized and normalized standard plasma matrix. The ratio of the coefficient of variation for most markers is much greater than one, so the variability across plates has been substantially reduced.

As shown in the tables 17 and 18, except a few ratios of c.v. are smaller but close to 1, most ratios are much greater than 1. Table 19 shows the mean ratios of c.v. between unnormalized and normalized Sample B matrix. The mean ratio for 21-marker is 2.19, and for 27-marker is 2.82. Comparing with the previous normalization methods, the mean ratios are the highest. Lowess curve extrapolation is the most effective solution to reduce the variability across plates. The lowess curve extrapolation is performed by R programming where span is set up as default value 0.75.

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|
| 1 | 0.90 | 12 | 3.43 |
| 2 | 1.53 | 13 | 1.42 |
| 3 | 1.11 | 14 | 3.27 |
| 4 | 2.08 | 15 | 1.51 |
| 5 | 1.39 | 16 | 2.02 |
| 6 | 1.32 | 17 | 1.67 |
| 7 | 2.00 | 18 | 1.50 |
| 8 | 2.76 | 19 | 1.61 |
| 9 | 1.89 | 20 | 4.76 |
| 10 | 1.08 | 21 | 6.32 |
| 11 | 2.39 | | |

**Table 20: Lowess curve extrapolation for the first 21 markers: ratio of coefficient of variation (CV) between unnormalized and normalized Sample B matrix**

| Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. | Marker ID | Ratio of C.V. |
|---|---|---|---|---|---|
| 1 | 0.70 | 10 | 0.71 | 19 | 0.70 |
| 2 | 0.97 | 11 | 0.73 | 20 | 0.92 |
| 3 | 0.92 | 12 | 1.04 | 21 | 0.70 |
| 4 | 0.82 | 13 | 0.80 | 22 | 0.57 |
| 5 | 0.96 | 14 | 0.68 | 23 | 0.72 |
| 6 | 2.79 | 15 | 0.69 | 24 | 0.70 |
| 7 | 1.01 | 16 | 0.87 | 25 | 0.69 |

| | | | | | |
|---|---|---|---|---|---|
| 8 | 0.84 | 17 | 0.70 | 26 | 0.84 |
| 9 | 0.79 | 18 | 1.07 | 27 | 0.67 |

**Table 21: Lowess curve extrapolation for the remaining 27 markers: ratio of coefficient of variation (CV) between unnormalized and normalized Sample B matrix**

| | mean ratios of c.v. |
|---|---|
| 21-marker, standard plasma | 0.69 |
| 27-marker, standard plasma | 0.87 |

**Table 22: mean ratios of c.v. between unnormalized and normalized Sample B matrix**

## 3.4 Evaluation of Normalization Methods

Since several normalization methods have been used on the Luminex assay system, which normalization method performing best needs to be assessed.

Spearman's rank correlation coefficient is a solution to assess how agreeable between two arrays if they are more or less, monotonically related to each other.

$$\hat{\rho}_s = \frac{12\sum_{g=1}^{G}\{R_{g1} - \frac{1}{2}(G+1)\}\{R_{g2} - \frac{1}{2}(G+1)\}}{G(G^2-1)}$$, where g = 1, 2, … G (number of genes); $R_{gi}$ is the rank of $y_{gi}$ when the $\{y_{gj}\}$ are ranked from 1 to G.

In general, Spearman's rank correlation coefficient is a measure of monotone association between two variables, which is not necessarily linear association. The range of correlation is between -1 and 1. When $\hat{\rho}_s$ is 1 the two set of values are perfectly positively associated to each other, and when $\hat{\rho}_s$ is -1 this indicates the two sets of values are negatively associated with each other, and 0 indicates no association associated with each other. Based on this concept, when normalizing two arrays, if $\hat{\rho}_s$ is high or close to1, it is like the normalization brings the two sets

of arrays into an agreement. When $\hat{\rho}_s$ is low it is unlikely that a normalization of that sort would be able to bring the two sets of arrays into agreement.

In addition, Lin et al[37] introduced concordance correlation coefficient that is used to assess the degree of success of the normalization. This is an index that quantifies the degree of agreement between two sets of numbers. The concordance correlation coefficient, $\hat{\rho}_c$, is defined as:

$$\hat{\rho}_c = \frac{2s_{12}}{s_1^2 + s_2^2 + (\overline{y_1} - \overline{y_2})^2}$$

Where $\overline{y_c} = \sum_{g=1}^{G} y_{gc} / G$, mean of the $i^{th}$ microarray

$S_c^2 = \sum_{g=1}^{G} (y_{gc} - \overline{y_c})^2 / G$, variance of the $i^{th}$ microarray.

$s_{12} = \sum_{g=1}^{G} (y_{g1} - \overline{y}_1) \cdot (y_{g2} - \overline{y}_2) / G$, the covariance between 2 microarrays

c = 1, 2 (2 arrays), g = 1, 2, … G (number of genes)

$\rho_c$ is a standardized measure of $E[(y_{g1} - y_{g2})^2]$ and $\rho_c = 1$ if and only if $\{y_{g1}\}$ and $\{y_{g2}\}$ are in perfect agreement. Otherwise $\rho_c < 1$.

Amaratunga et.al [4] in their book proposed the rough rule of thumb to evaluate the need for normalization, when combine Spearman's rank correlation coefficients and concordance correlation coefficients together. 1) For a pair of arrays, if $\hat{\rho}_c$ is very high (a rule of thumb is greater than 0.99), normalization may not be necessary. 2) if $\hat{\rho}_c$ is not very high and $\hat{\rho}_s$ is high( a rule of thumb is greater than 0.8), normalization is very likely to be beneficial. 3) If both $\hat{\rho}_c$ and $\hat{\rho}_s$ are low, indicating the relationship between arrays is not strong, it may be worth looking further to see whether there was a problem with either of arrays before doing any normalization.

For a series of arrays, Amaratunga et al [4] also suggests to display the pair-wise Spearman's rank correlation coefficients or pair-wise concordance correlation coefficients. However, the correlation coefficient is not appropriate to be applied to this project. This is because within each plate the scales across markers are quite different. Most markers are expressed under 200, some are expressed over 1000. So even the two arrays are in disagreement and not correlated with each other at all, the outliers (some intensity over 1000) could potentially elevate the correlation to be close to 1. But this is not true that we could regard two arrays are in quite agreement.

Table 19 shows the mean of each marker (for marker27) across Sample A matrix. We notice that most markers are scaled between 30 and 300. However marker 24 and 25 stand out with quite different scales. For marker 24, its mean is 596.3875 and the range is: (489.3, 757.5). For marker 25, its mean is 3938.863 and the range is: (3645.5, 4295.2).

| Marker ID | Mean | Marker ID | Mean | Marker ID | Mean |
|---|---|---|---|---|---|
| 1 | 232.95 | 10 | 55.36 | 19 | 52.49 |
| 2 | 42.84 | 11 | 83.03 | 20 | 47.70 |
| 3 | 38.66 | 12 | 69.73 | 21 | 302.80 |
| 4 | 36.31 | 13 | 42.43 | 22 | 164.95 |
| 5 | 56.89 | 14 | 48.26 | 23 | 71.55 |
| 6 | 37.48 | 15 | 64.36 | 24 | 596.39 |
| 7 | 50.18 | 16 | 191.09 | 25 | 3938.86 |
| 8 | 63.70 | 17 | 67.35 | 26 | 68.08 |
| 9 | 74.35 | 18 | 78.68 | 27 | 101.66 |

**Table 23: mean of markers (marker-27) from Sample A matrix**

Table 20 shows the pair-wise concordance correlation coefficient among 8 plates in 27-marker system. There are very high correlations between plates that are almost 1. Without considering the fact these markers are quite different scales, it is misleading to conclude that plates are very concordant with each other thus normalization might not be necessary.

| Marker | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | 0.993 | 0.991 | 0.982 | 0.990 | 0.997 | 0.998 | 0.991 |
| 2 | 0.993 | 1.00 | 0.999 | 0.997 | 0.999 | 0.998 | 0.998 | 0.999 |
| 3 | 0.991 | 0.999 | 1.00 | 0.998 | 0.999 | 0.996 | 0.997 | 0.999 |
| 4 | 0.982 | 0.997 | 0.998 | 1.00 | 0.998 | 0.991 | 0.991 | 0.998 |
| 5 | 0.990 | 0.999 | 0.999 | 0.998 | 1.00 | 0.995 | 0.997 | 0.999 |
| 6 | 0.997 | 0.998 | 0.996 | 0.991 | 0.995 | 1.00 | 0.999 | 0.996 |
| 7 | 0.998 | 0.998 | 0.997 | 0.991 | 0.997 | 0.999 | 1.00 | 0.997 |
| 8 | 0.991 | 0.999 | 0.999 | 0.998 | 0.999 | 0.996 | 0.997 | 1.00 |

**Table 24: pair-wise concordant correlation coefficient among plates (27-marker)**

Table 21 shows the concordance correlation coefficient after we remove the marker 24 & 25 from the control plasma matrix. Therefore the coefficients correlation went down and the two arrays do not "seem" to in a perfect match. This demonstrates that when different scales of intensities exist in the dataset, utilizing concordance correlation coefficient to assess the performance of normalization is not appropriate.

| Marker | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | 0.867 | 0.930 | 0.832 | 0.923 | 0.871 | 0.974 | 0.923 |
| 2 | 0.867 | 1.00 | 0.963 | 0.976 | 0.978 | 0.985 | 0.909 | 0.975 |
| 3 | 0.930 | 0.963 | 1.00 | 0.958 | 0.977 | 0.958 | 0.970 | 0.977 |
| 4 | 0.832 | 0.976 | 0.958 | 1.00 | 0.945 | 0.965 | 0.887 | 0.955 |
| 5 | 0.923 | 0.978 | 0.977 | 0.945 | 1.00 | 0.970 | 0.961 | 0.989 |
| 6 | 0.871 | 0.985 | 0.958 | 0.965 | 0.970 | 1.00 | 0.902 | 0.983 |
| 7 | 0.974 | 0.909 | 0.970 | 0.887 | 0.961 | 0.902 | 1.00 | 0.952 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 8 | 0.923 | 0.975 | 0.977 | 0.955 | 0.989 | 0.983 | 0.952 | 1.00 |

**Table 25: pair-wise concordant correlation coefficient among plates (27-marker) after marker 24 and 25 are removed**

Park et al. [38] proposed variability among the replicated slides to compare performance of normalization methods. They introduced variation $\sigma_l^2$ , l = 1, 2, …N genes in a slide to evaluate the success of a normalization, and proposed two methods for estimating $\sigma_l^2$ .

The first method is pooled variance estimators: for gene l, a simple variance estimator for $\sigma_l^2$ is estimated as:

$$\hat{\sigma}_l^2 = \frac{1}{IJ(k-1)} \sum_i \sum_j \sum_k (y_{ijkl} - \bar{y}_{ij.l})^2$$

Where $\bar{y}_{ij.l} = \frac{1}{k} \sum_{k=1}^{K} y_{ijkl}$

$\bar{y}_{ij.l}$ is the intensity from experimental groups i (i =1, …, I), j time point, and k (=1, …, K) replications and gene l.

The second method is variance estimator using analysis of variance models: consider the following two-way analysis of variance (ANOVA) model with interactions for each gene:

$$y_{ijkl} = \mu_l + \alpha_{il} + \beta_{jl} + (\alpha\beta)_{ijl} + \varepsilon_{ijkl}$$

Where i = 1,…, I, j = 1, …, J, k=1, …, K and l = 1, … N. $\mu_l$ is the gene effects that capture the overall mean intensity across arrays, groups and time points. $\alpha_{il}$ accounts for gene specific group effects representing overall difference between two groups. $\beta_{jl}$ is the time effects that capture difference in the overall concentration of mRNA in the samples from the different time points. $(\alpha\beta)_{ijl}$ is the interaction effect between group and time. From this ANOVA model, the unbiased estimate of variance for the $l^{th}$ gene $\sigma_l^2$ is obtained by error sum of squares divided by

degrees of freedom. If there are no missing observations, the variances estimated from ANOVA are the same with pooled variance. However, ANOVA gives more flexibility to fit the intensity data.

Some people also suggest using bias and mean square error as well as variance to compare between normalization methods. Bolstad et al [14] used variance and bias to compare normalization method for high density oligonucleotide array data. Likewise, for each gene or marker, the mean square error (MSE) is computed as the average of the distances between normalized data and true expected value. However, in a real dataset, we don't know true values so we cannot calculate bias or MSE.

In this project, we use the coefficient of variation, which is the ratio of standard deviation to the absolute mean, across the control plasma matrix / standard plasma matrix to compare the performance of normalization methods. It is expected that the better the normalization method, the smaller the coefficient of variation. The coefficient of variation in each marker across control/standard plasma matrix is expressed as $\text{c.}v._{i,}\ i = 1, 2, \cdots N$ . Dot plot for the variability measures can be used for visually comparing different normalization methods.

We denote $y_{ij}$ is the intensity for plate i (i =1,2, …8) and marker j (j =1, 2, …,48), $\text{c.}v_j$ is the coefficient of variation for the $j^{th}$ marker. The coefficient of variation for $j^{th}$ marker is:

$\text{c.}v._j = \frac{\sigma_j}{|\mu_j|}$, where $\sigma_j$ is the standard deviation, and $\mu_j$ is the mean for the $j^{th}$ marker across 8 plates. We compared the ratio of coefficient of variation between unnormalized and normalized standard plasma matrix. The comparison is based on four methods: scale (median) normalization, quantile normalization, lowess curve normalization, and lowess curve extrapolation. Specifically,

we compare the degree to which each method reduces the coefficient of variation across plates for Sample A ("normal" control) and Sample B (combination of normal control and Standard 6).

## 3.5 Discussion

Undesirable systematic variations are commonly observed in the microarray experiment. Normalization becomes a standard process for removing some of the variation which affects the measured gene expression levels.

In this dissertation, we compared normalization methods commonly used to analyze microarray data. The comparison is based on the variability reduction from control plasma or standard plasma matrix, which are used to evaluate the performance of normalization.

Although a number of normalization methods have been proposed, it has been difficult to decide which method performs better than the others. Therefore the evaluation of normalization methods in microarray data analysis is indeed an important issue. We used coefficient of variation to evaluate performance of each method.

Figure 12 and 13 are the dot plots that compare the ratios of c.v. among four normalization methods. We expect the higher the ratios between unnormalized and normalized, the more variation reduced.

In 21-marker, ratios of c.v. between unnormalized and normalized Sample B matrix scatter around 1 for both median and lowess curve normalization. For quantile normalization, most ratios are around 1 but a few are below 1. Lowess curve extrapolation outperforms other three methods, as shown its ratios are scatter above y =1.

For 27-marker, on the average ratios in both median and lowess curve normalization methods fluctuate around 1. However, quantile normalization has most ratios below than 1. Lowess curve extrapolation have ratios that are much higher than 1. There is substantial reduction in lowess

curve extrapolation. Table 23 and 24 compare the mean ratios among four methods in 21-marker and 27-marker respectively.

In conclusion, for this dataset, lowess curve extrapolation reduces the variation the most, followed by median normalization & lowess curve normalization, then quantile normalization. Scale normalization is the simplest approach assuming average gene expression is the same for all arrays, and most genes change very little in intensity across samples. This is justified since the equal quantities of mRNA were distributed to the samples, therefore the average hybridization should be the same for all samples.

Quantile normalization does not work particularly well for this dataset. By its definition, when quantile normalization is applied on thousands of genes at the same time, it guarantees that the normalized intensities distribution is the same on each plate. However in our current dataset, each plate includes either 27 or 21 markers. The sample size is not sufficient enough to perform the quantile normalization well to reduce variation.

Although we have studies a limited number of normalization, our findings can provide some guidance on the selection of normalization methods. We think the non-linear normalization methods such as lowess curve extrapolation is quite effective in controlling specific non-linear variations. Some methods may have more computational efficiency. Please note that the complex methods do not necessarily perform better than simpler methods. Complex methods may add noise to the normalized adjustment and may even add bias if the assumptions are not justified. Consequently the complicated normalization methods require validation. We suggest researchers examine their data carefully and consider applying non-linear normalization routinely. The normalizations are performed by R programming and available upon the request.

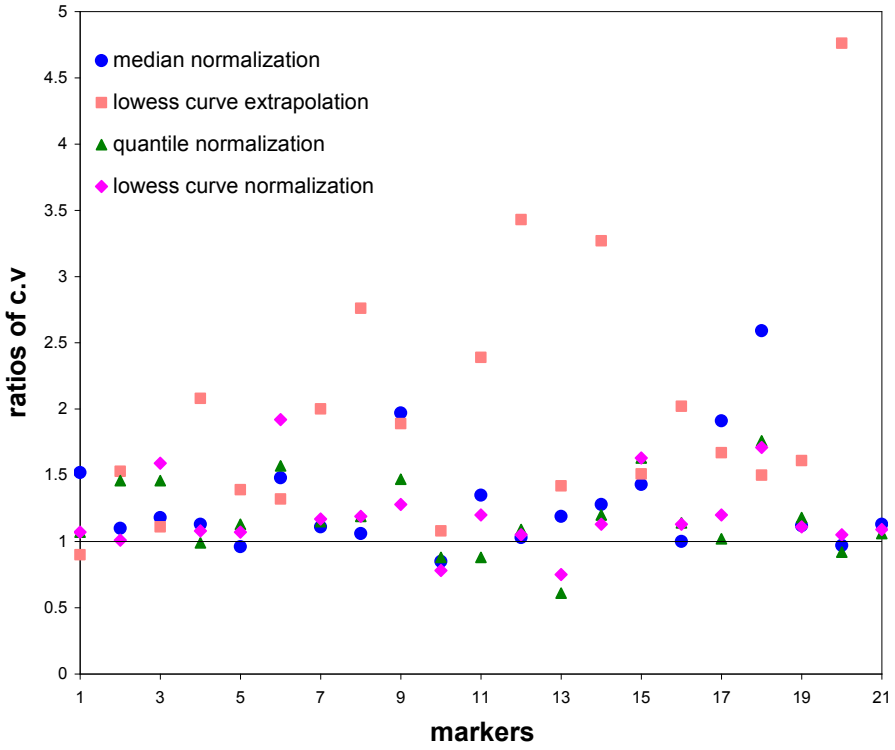


**Figure 13: Dot plots of the ratios of coefficient of variation between unnormalized and normalized Sample B matrix. The y-axis is the ratios of coefficient of variation, and the x-axis is each marker in Sample B matrix (21-markers)**

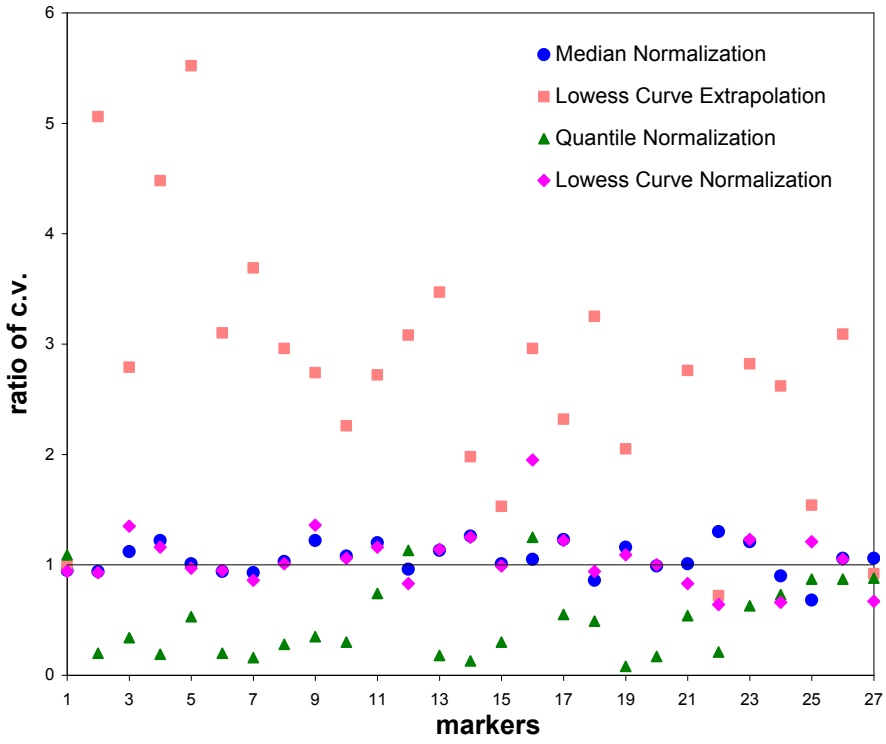

**Figure 14: Dot plots of the ratios of coefficient of variation between unnormalized and normalized Sample B matrix. The y-axis is the ratios of coefficient of variation, and the x-axis is each marker in Sample B matrix (27-markers)**

| Normalizations | Plex | Sample A | Sample B |
|---|---|---|---|
| Median | 21 | 1.02(0.02) | 1.30(0.41) |
| | 27 | 1.23(0.27) | 1.05(0.14) |
| Quantile | 21 | 1.08 (0.43) | 1.18 (0.28) |
| | 27 | 0.51(0.52) | 0.50(0.34) |
| Lowess | 21 | 1.14(0.44) | 1.20(0.29) |
| | 27 | 0.92(0.41) | 1.05(0.26) |
| Lowess Curve Extrapolation | 21 | 0.69 (0.23) | 0.87 (0.19) |
| | 27 | 0.93(0.44) | 0.95 (0.21) |

**Table 26: Comparison of normalization methods: mean ratios of unnormalized to normalized c.v.**

# Chapter 4: Evaluation of methods on a second example

We evaluated the three normalization methods using data from a second study. This is an observational study "biologic changes in lung transplant recipients". The 1088 samples from 309 transplant recipients are bronchoalveolar lavage fluid collected during surveillance and clinically indicated bronchoscopies from lung transplant recipients at UCLA between July 2000 and June 2008. The study was to explore the inflammatory milieu in the lung after transplantation, and how alterations in the milieu could inform common pathologies post-transplantation (i.e. acute rejection, chronic rejection, infections, etc.). The panel was a standard human inflammation/immunology multiplex panel from Millipore and the assay was done upon the Luminex assay system.

Samples were run on a total of 12 plates each with 42 markers and fluorescent intensities were obtained. On each plate control 1 and 2 were spiked samples provided by the manufactures with known values for each analyte and they were used to estimate the percent recovery. In plate 1 controls were duplicated. Two samples c1 and c2 with unknown values were run from plate 2 to 12. Since controls and samples were run on most plates, we could use them to infer plate to plate variance. Therefore we have four plasma matrices to evaluate how normalization methods have reduced the plate to plate variance. Control 1 or 2 run from plate1 to plate 12, and they could be organized as two matrices where the dimension is 13*42. "13" is for 12 plates where plate1 had duplicates for controls. "42" is for 42 markers. Sample 304 or 530 run from plate 2 to 12 and they were organized as another two matrices where the dimension is 11*42.

Results are shown in Table 6 and Figure 4 for control 1 and 2, samples 1 and 2.

| | Average ratio of c.v. across plates | | | |
|---|---|---|---|---|
| | Control 1 | Control 2 | Sample 1 | Sample 2 |
| Median normalization | 1.15(0.12) | 1.52(0.37) | 1.17(0.17) | 1.14(0.20) |
| Quantile normalization | 0.95(0.27) | 0.97(0.14) | 0.96(0.30) | 0.95(0.30) |
| Lowess curve normalization | 0.99(0.26) | 1.01(0.12) | 1.27(0.29) | 1.26(0.23) |

*mean ratios of coefficient of variation are followed by standard deviation

**Table 27: Comparison of normalization methods**

Normalization results from control or sample matrices are consistent with our findings from the lung cancer dataset. Median normalization reduces variation across plates the most, followed by the lowess curve normalization. Quantile normalization is not steady in reducing the variation across plates for each marker. Figure 14 illustrates that median normalization gives higher ratios of coefficient of variation than other two methods, based on the four assessment matrices.

In addition, we have investigated the relationship between ratios of coefficient of variation and the scale of each marker, and we don't find a particular pattern between them. The ratios fluctuate around 1 across markers. The ratios of the coefficient of variation are independent of the intensity in each marker.

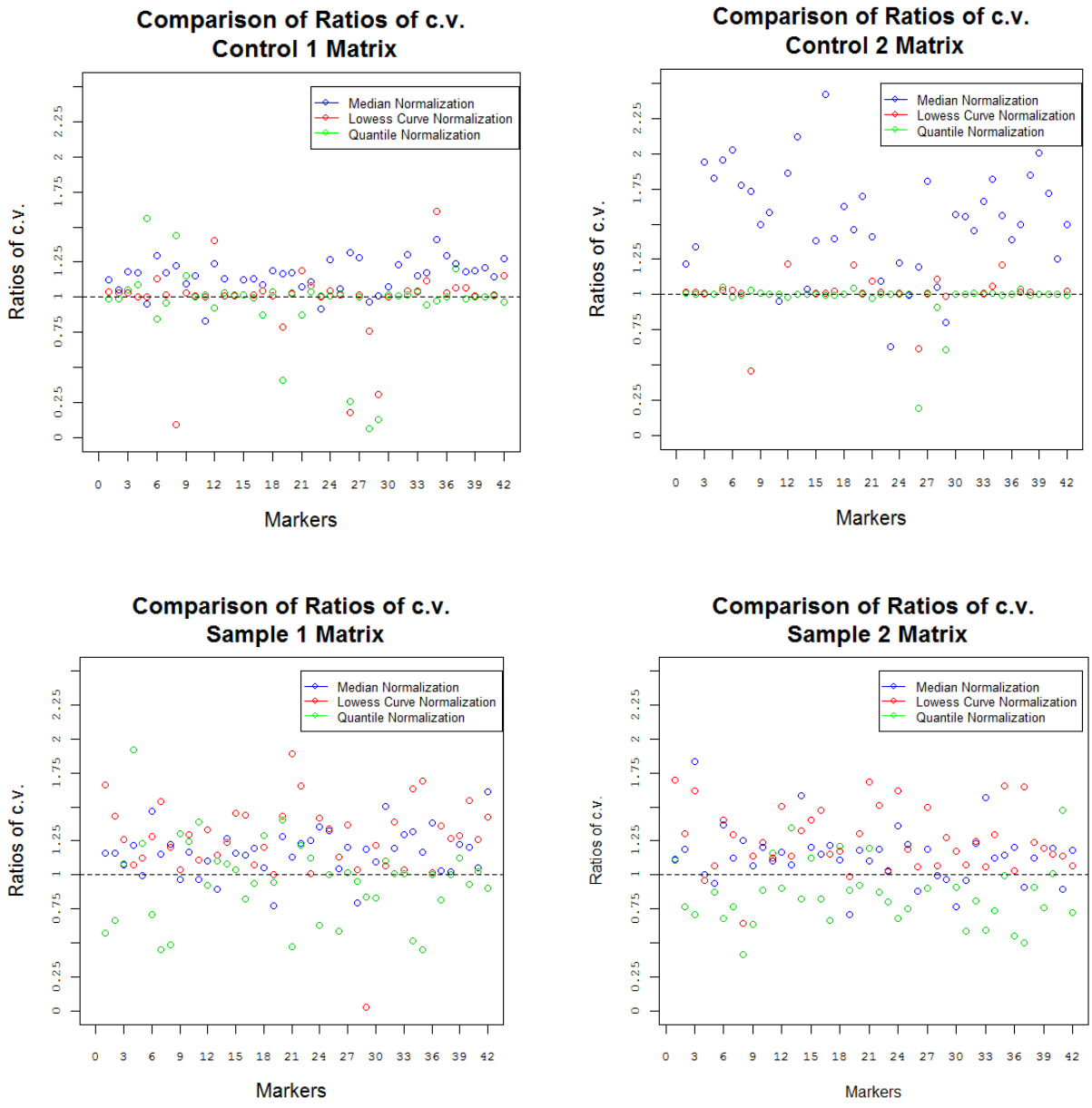


**Figure 15: Comparison of normalization methods on control 1 and 2, and samples c1 and c2.**

# Chapter 5: Integration of miRNA and mRNA datasets

## 5.1 Introduction

MicroRNAs (miRNAs) are small non-coding, endogenous, 19-24nt single-stranded RNAs acting as post-transcriptional regulators of gene expression [39-40].The first 2 miRNAs, lin-4 and let-7, were experimentally discovered in 1993 and 2000[41-42]. So far there are more than 4300 miRNAs that have been identified in plants, animals, and viruses by cDNA sequencing and computational predictions [43-45]. miRNA regulate mRNA and protein levels through base-pairing, by inducing mRNA degradation and translational repression[46-47]. By modulating key cellular processes such as metabolism, division, differentiation, they can simultaneously regulate both oncogenes and tumor suppressor genes [48-49]. Deregulation of miRNAs expression plays a critical role in the pathogenesis of genetic and multifactorial disorders, as well as most human cancers[49]. According to the increasing experimental evidences supporting the miRNA mechanism of target degradation, miRNAs tend to negatively regulate mRNAs, i.e. the expression profiles are expected to be anti-correlated.

There are two ways that miRNA integrates mRNAs depending on the degree of complementarity between miRNA and its targets [50]. One is miRNAs bind perfectly to their targets' coding sequence and they are thought to result in mRNA degradation. Another one is miRNAs bind with imperfect complementarity to the 3' UTR block target gene expression at the level of protein translation. Given that miRNAs can have multiple targets and that each protein-coding gene can be targeted by multiple miRNAs, it has been suggested that more than 1/3 human genes could be regulated by miRNAs[51].

For target prediction, the database including MiRecords[52] and TarBase[53], are becoming more extensive. However bioinformatic algorithms remain the principal means of predicting targets of specific miRNAs. Currently three of the most widely used and predicatively accurate algorithms are TargetScan, PITA, and PicTar. Two other methods Miranda and mirWIP are also increasingly being used[46]. For the webtool, miRGator[54] and DIANA-microT[55] web servers are used to elucidate the biological processes, functions and pathways targeted by miRNAs. The webtool MMIA[40] (miRNA and mRNA Integrated Analysis) integrates miRNAs and mRNA expression data using significantly up or down-regulated features, but it does not take into account the whole expression profile and loses the key information for the calculation of the expression anti-correlation degree. MAGIA [56] (miRNA and gene integrated analysis) is another webtool that integrates target predictions and gene expression profiles using miRNA-mRNA bipartite networks reconstruction, gene functional enrichment and pathway annotations.

In conclusion, genome-wide miRNA studies allow the investigation of genomic changes at the miRNA level and are likely to provide additional clues to the mechanisms of tumorigenesis. Particularly, when miRNA and mRNA expression are both measured on the same samples, an integrative analysis can be performed to compare miRNAs and mRNAs profiles and to study their interaction patterns. We will perform the integrative analysis using a published study of miRNA and mRNA expression in renal cell carcinoma samples in section 5.2.

## 5.2 Study Illustrations

We used expression data from clear cell Renal Cell Carcinoma (ccRCC) and matched normal kidney samples. In total RNA has sample size of 34, from 17 Renal Cell Carcinoma (RCC) tumors and 17 corresponding non-tumor samples. They were hybridized against a common reference RNA (Agilent-014850 Whole Human Genome Microarray) for gene

expression analysis. Correspondingly, MicroRNA from 17 RCC tumors and 17 corresponding non-tumor samples were hybridized on a single channel platform (Agilent Human miRNA Microarray Rel12.0) for miRNA expression analysis.

Renal Cell Carcinoma represents 3% of all malignancies in the US, with 50,000 new cases and 12,000 deaths each year. The most common histological class is ccRCC, which accounts for 75% of kidney cancers. ccRCC is known to be characterized by the loss of the VHL gene, where Von Hippel-Lindau syndrome (VHL) is a dominantly inherited familial cancer syndrome predisposing to a variety of malignant and benign tumors. Under normal oxygen pressure, ccRCC binds to the α subunits of hypoxia-inducible factors (HIFs), and induces subsequent degradation in the proteasome [57-58]. ccRCC tumors have a wide range of natural histories and varied responses to VEGF-targeted therapy. In early stage, low grade tumors tend to have significantly better disease free survival after resection than higher stage and grade [59]. Although VHL mutation is associated with all grades of ccRCC, the other molecular factors associated with ccRCC initiation and progressions are unknown. In conclusion, ccRCC is a ripe target for studies investigating the molecular and genetic nature of these hetero-genetics.

In RCC, various studies have identified panels of miRNAs and mRNAs that are differentially expressed between normal renal tissue and tumor [60-64]. Liu et al.[65] linked the miRNA to some of their putative gene targets, thus uncovering an unknown part of the biology of ccRCC disease. They also identified miRNA/mRNA anti-correlation relationship and validated this on a new cohort ccRCC study.

## 5.3 Normalization of miRNA and mRNA

The study includes 33398 mRNAs. The raw dataset has been log 2 transformed and lowess normalized. We first related gene expression with their gene symbols and entrezene, and

calculated the mean expression for RCC (renal carcinoma, 17 sample size) and normal tissues (17 sample size).

There are 319 miRNA in the study. To comply with the normalization of mRNA, we log 2 transformed miRNA and performed quanitile normalization on it. After quantile normalization, the expression on each sample is matched with each other. For instance, Figures 15 and 16 illustrated the pair-wise comparison between RCC sample 1 with sample 2, 3, 4, 5 before and quantile normalization.

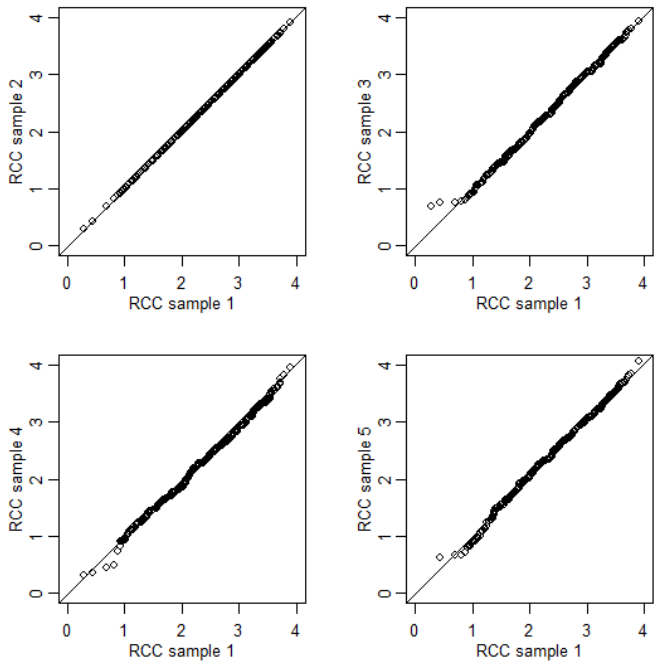


**Figure 16: Pair-wise comparison between RCC sample 1 and sample 2, 3, 4, 5 before quantile normalization**

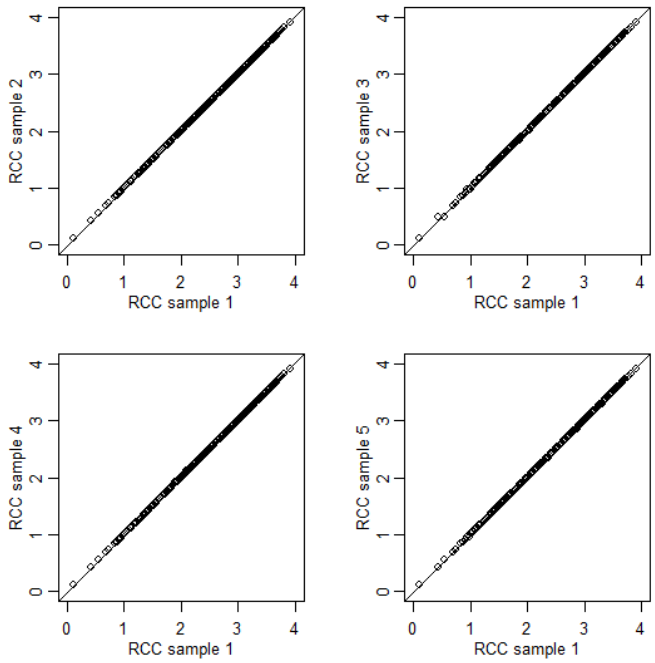


**Figure 17: pair-wise comparison between RCC sample 1 and sample 2, 3, 4, 5 after quantile normalization**

The underlying hypothesis in our method is that the expression levels of miRNAs and their putative genes are strongly correlated when averaged over matched samples in either tumor or normal tissue.

The stepwise procedure is as follows:

1) Log 2 transformed and normalized the miRNA and mRNA.
2) For normalized mRNAs, using WGCNA package to perform step-by-step network construction and module detection. Thus we partitioned highly correlated gene into clusters (referred to as groups or modules).
3) Identify strong correlation between each miRNA and module eigengenes. Eigengenes summarizes the overall behavior of a module. We computed the Pearson correlation between miRNA and module eigengenes and retain only those whose correlation $|\rho| >= 0.7$.

4) For each of 319 miRNAs, using "TargetScan" to look for its putative gene targets. "TargetScan" uses seed sequence complementarity and free energy predictions of RNA-RNA duplexes to identify the putative targets [53, 66-72]. We found that out of 319 miRNAs, 257 miRNAs could be located by their putative targets through "TargetScan".

5) We next evaluate for each miRNA, whether its putative genes are differentially present between a given module and other modules by the Fisher's exact test. In the 2 by 2 contingency table of the Fisher's exact test, row is if miRNAs belong to a given module or not, and column is if these miRNAs are putative targets of a specific miRNA or not. We are particularly interested in the 1$^{st}$ cell that belongs to a given module and putative gene targets. The Fisher's exact test hypothesize that putative gene targets have the same patterns between a given module and other modules. If the Fisher's exact test is significant, it suggests there is a different pattern between a given module and other modules in terms of putative miRNAs. For this module, it has higher frequencies of putative genes than we expect by chance alone.
6) Retain only those miRNAs who are strongly correlated with their putative genes targets and also significant in the Fisher's exact test.

## 5.4 Clustering Samples

Clustering samples is a distance measure which is calculated between the expression profiles of each gene (or clusters/modules) pair, and a recursive bottom-up or top-down algorithm to merge or split genes based on their distance. Examples are Euclidean distance and one minus the Pearson correlation coefficient. Hierarchical clustering is the most commonly used method for samples clustering using expression profiles[73].

A common drawback of clustering is it always generates a clustering even when there is no real underlying clustering in the dataset. It is not apparent whether the clustering structure

reflects a true pattern in the data or just an artifact of the clustering algorithm. Methods based on resampling simulate perturbations of the original data and assess the stability of the clustering results [74-76].

We will introduce a few clustering methods run by WGCNA package in the following section. The method applied to our project is hierarchical clustering to identify stable samples clusters (modules) based on their gene expression.

## 5.5 WGCNA package: Step-by-step Network Construction and Module Detection

Correlation networks are increasingly being used in bioinformatics applications. For example, weighted gene co-expression network analysis (WGCNA) is a systems biology method for describing the correlation patterns among genes across microarray samples. It can be used for finding clusters (modules) of highly correlated genes, for summarizing such clusters using the module eigengene or an intra-modular hub gene, for relating modules to external sample traits by using eigengene network methodology. Correlation networks facilitate network based gene screening methods that can be used to identify candidate biomarkers or therapeutic targets. These methods have been successfully applied in various biological contexts, e.g. cancer, mouse genetics, and analysis of brain imaging data.[77].

The WGCNA software package is a comprehensive collection of R functions for performing various aspects of weighted correlation network analysis. The package includes functions for network construction, module detection, gene selection, calculations of topological properties,

etc. Here we used the function of step-by-step network construction and module detection to find clusters of highly correlated genes.

WGCNA analyze genes by the following steps:

1) Assessing scale free topology and choosing the parameters of the adjacency function using the scale free topology criterion (Zhang and Horvath[78])
2) Computing the topological overlap matrix
3) Defining gene modules using clustering procedures
4) Summing up modules by their first principal component (first eigengene)
5) Relating a measure of gene significance to the modules
6) Carrying out a within module analysis (computing intramodular connectivity) and relating intramodular connectivity to gene significance.
7) Miscellaneous other functions, e.g. for computing the cluster coefficient.

Detecting clusters (modules) of closely related genes is an important step in genetics. The WGCNA package uses dissimilarity measure matrix $D= (d_{ij})$ to assign the cluster. The matrix D is symmetric whose diagonal elements equal to 0. There are a few concepts we shall introduce first.

Total object scatter:

$$TotalScatter(D) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} = \sum_{i=1}^{n}\sum_{j>1}^{n} d_{ij}$$

Within object scatter:

$$WithinScatter(Cl, D) = \frac{1}{2}\sum_{q=1}^{k}\sum_{Cl(i)=q}\sum_{Cl(j)=q} d_{ij}$$

Between cluster object scatter:

$$BetweenScatter(Cl, D) = \frac{1}{2}\sum_{q=1}^{k}\sum_{Cl(i)=q}\sum_{Cl(j)\neq q} d_{ij}$$

To define the cluster assignment Cl, we need to look for the Cl that minimizes the within cluster object scatter WithinScatter (Cl, D) over all possible assignments of the n objects to k clusters.

Mainly there are three clustering methods that are widely used in network application. They are as follows:

1) Partitioning-around-medoids (PAM) clustering: this is a clustering procedure implementing an iterative algorithm for minimizing the within-cluster scatter WithinScatter (CL). A medoid is the most centrally located object inside a given cluster. PAM is a classical partitioning technique to cluster a set of data of n objects into k clusters[79]. k is the number of cluster where we need to specify first.

The PAM algorithm is as follows:

Step 1: The algorithm begins with k medoids where k is pre-specified.

Step 2: Each of the remaining objects are assigned to the medoid that is least dissimilar. This result in a cluster assignment CL. Compute the within-cluster scatter WithinScaller(CL).

Step 3: For each of the k cluster, determine the medoid i.e. the object which minimizes the average dissimilarity to the other objects in the cluster.

Step 4: Compute the within cluster scatter $WithinScatter_{newmedoids}(Cl)$ after swapping the initial medoids with the new set of medoids.

Step 5: If $WithinScatter_{newmedoids}(Cl) < WithinScatter(Cl)$, then swap the initial set of k medoids with the new set.

Repeat the steps 2 to 5 till no change in the medoid assignments.

2) Agglomerative Hierarchical Clustering:

It begins objects as a separate cluster and merges them into successively larger clusters. Hierarchical clustering creates clusters that are represented in a tree structure (dendrogram). The root of the tree consists of a single cluster containing all objects, and the leaves correspond to individual objects.

Agglomerative hierarchical clustering has 2 inputs: 1) a pair-wise dissimilarity measure. 2) inter-cluster dissimilarity which is based on the pair-wise dissimilarities between objects inside the clusters. There are 3 approaches to define the inter-cluster dissimilarity between clusters $clust.q1$ and $clust.q2$.

The average linkage hierarchical clustering:

$$d_{average}(clust.q1, clust.q2) = \frac{\sum_{i \epsilon clust.q1} \sum_{j \epsilon clust.q2} d_{i,j}}{|clust.q1||clust.q2|}$$

Complete linkage clustering:

$$d_{complete}(clust.q1, clust.q2) = \max\left(\left\{d_{i,j} \middle| i\epsilon\ clust.q1, j\epsilon\ clust.q2\right\}\right)$$

Single linkage clustering:

$$d_{single}(clust.q1, clust.q2) = \min\left(\left\{d_{i,j} \middle| i\ \epsilon\ clust.q1, j\ \epsilon\ clust.q2\right\}\right)$$

In our project, we used the agglomerative hierarchical clustering and average linkage clustering, and it leads to robust clusters. The R function involved is flashClust that implements the algorithm of order $n^2$ (n is the number of clustered objects).

3) DynamicTreeCut method and R package

WGCNA package developed a "dynamic" branch cutting method based on analyzing the shape of the branches, as opposed to static. The algorithm is implemented in the

DynamicTreeCut R package[80]. There are two variants of the method. The first variant, invoked function cutreeDynamicTree, is a top-down algorithm that only input the cluster tree. The second variant, invoked using function cutreeHybrid, is a bottom-up algorithm that inputs both a cluster tree and dissimilarity measure. It is a hybrid between hierarchical clustering and partitioning-around-medoids (PAM) clustering, and it improve the detection of outlying members of each cluster.

To implement WGCNA package, we first chose the soft thresholding power β to which co-expression similarity is raised to calculate adjacency[78]. Zhang and Horvath proposed to choose the soft thresholding power based on the criterion of approximate scale-free topology. The criteria are to choose the lowest β that results in approximate scale free topology as measured by the scale free topology fitting index.

$$R^2 = ScaleFreeFit = cor(\log(p(dk)), \log(BinNo))^2$$

WGCNA package uses the function **pickSoftThreshold** that performs the analysis of network topology and aids the user in choosing a proper soft-thresholding power. β = 6 is the default choice for unsigned weighted networks. The results are shown in Figure 17. On the left side the Scale Free Topology index $R^2$ is a function of different powers β. When $R^2$ tend to go up with higher powers, there is not a strictly monotonic relationship. Instead, when the power = 20 the curve first reaches a saturation point. On the right side, the mean connectivity is strictly decreasing function of the power β. We chose power β =20 and it has $R^2$ 0.86 that is close to be 0.9. The advantage of weighted network is that they are highly robust with regard to the power β, i.e. other choices also lead to similar modules. Notice that there is a trade-off between

maximizing the scale-free topology model fit and maintaining a high mean number of connectivity. Power that leads to high $R^2$ tends to lead a network with very few connections. We also validate the choice of power β =20 from the free topology plot (Figure 18) using the functions **softConnectivity**. On the free topology plot the slope of the regression line between $\log_{10} P(k)$ and $\log_{10}(k)$ is around -1. This shows that the connectivity really follows the scale-free law.

| Power | Scale Free Topology $R^2$ | Mean Connectivity |
|---|---|---|
| 1 | 0.31 | 4712.49 |
| 2 | 0.02 | 1881.79 |
| 3 | 0.34 | 909.15 |
| 4 | 0.53 | 494.02 |
| 5 | 0.61 | 291.25 |
| 6 | 0.65 | 182.45 |
| 7 | 0.68 | 119.85 |
| 8 | 0.71 | 81.79 |
| 9 | 0.74 | 57.62 |
| 10 | 0.79 | 41.7 |
| 12 | 0.86 | 23.34 |
| 14 | 0.88 | 14.03 |
| 16 | 0.89 | 8.92 |
| 18 | 0.90 | 5.94 |
| 20 | 0.90 | 4.11 |
| 22 | 0.91 | 2.95 |
| 24 | 0.91 | 2.17 |
| 26 | 0.91 | 1.64 |
| 28 | 0.91 | 1.26 |
| 30 | 0.92 | 0.99 |

**Table 28: list of scale free topology under different powers**

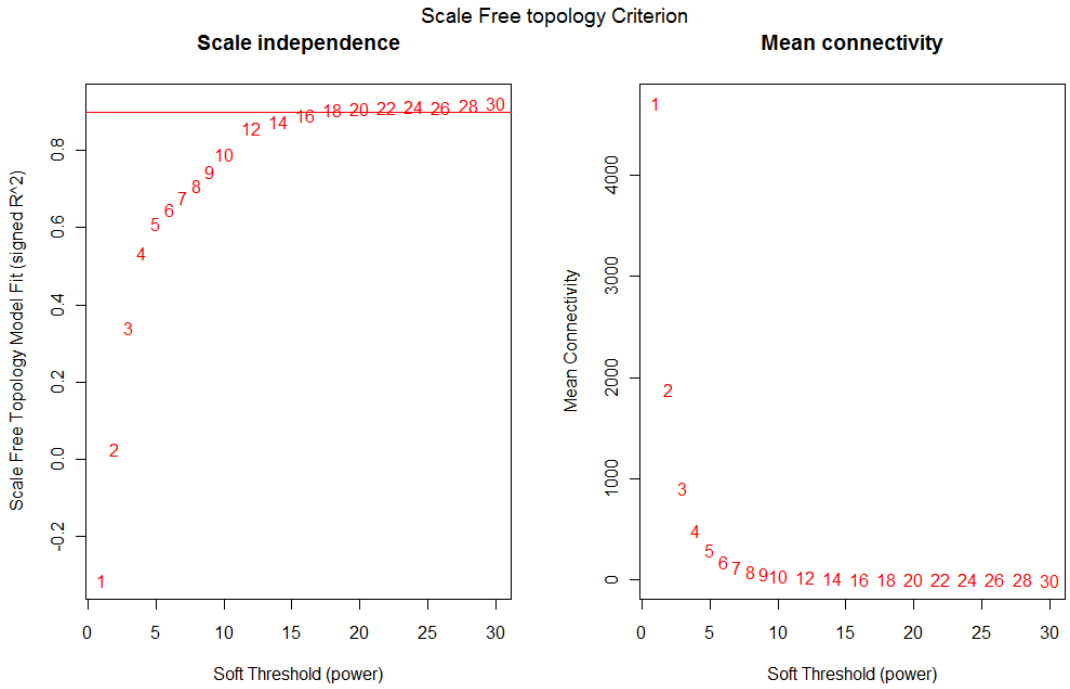


**Figure 18: Scale free topology for choosing the power β for the unsigned weighted correlation network.**

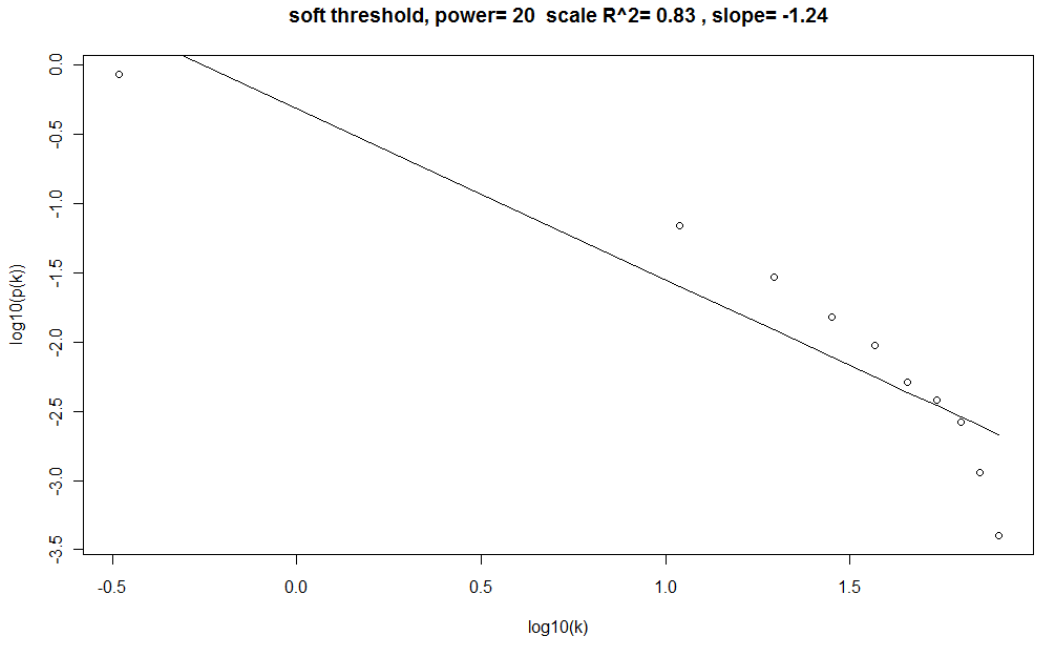


**Figure 19: The free topology plot shows the slope of the regression line between $\log_{10}$ P(k) and $\log_{10}$(k) is around -1.**

Then to minimize effects of noise and spurious associations, we transformed the adjacency into Topological Overlap Matrix (TOM) and calculated the topological overlap matrix based on the corresponding dissimilarity.

$$dissTOM_{ij} = 1 - TopOverlap_{ij}$$

Following the TOM, we used hierarchical clustering to produce a hierarchical clustering tree (dendrogram) of genes. The R function we have used is called **flashClust** which a fast hierarchical clustering routine. The branch cutting we used is the dynamic tree cut from the package **dynamicTreeCut**. We then defined modules as branches of the trees, i.e. module detection involves cutting the branches of the tree. Figure 19 demonstrates the clustering dendrogram. Notice that in the dendrogram, each vertical line corresponds to a gene. Branches of the dendrogram group densely interconnected and highly co-expressed genes together. Module identification amounts to the identification of individual branches.

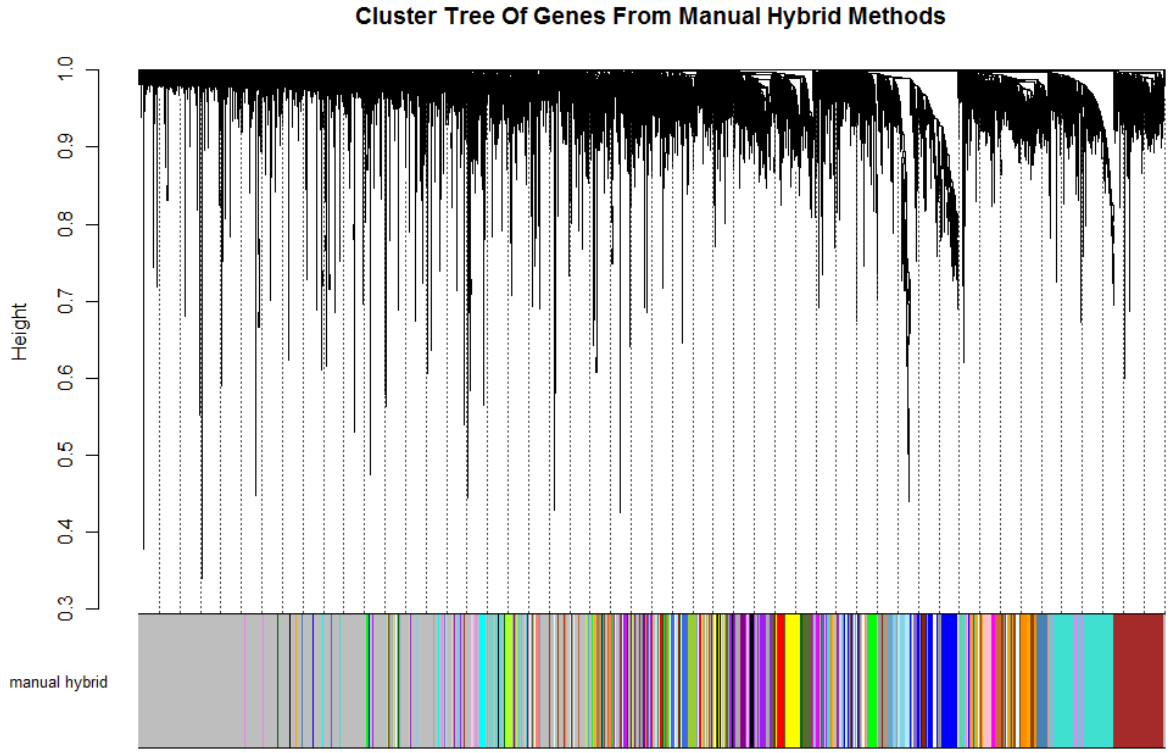

**Figure 20: Dendrogram before correlated modules are merged**

However some modules are very similar whose eigengenes are highly correlated. More specifically, Figure 20 visualizes the eigenegenes network representing the relationship among modules. The top panel shows a hierarchical clustering dendrogram of the eigengenes based on the dissimilarity $diss(q_1, q_2) = 1 - cor\left(E^{(q_1)}, E^{(q_2)}\right)$. The bottom panel shows the eigengenes adjacency $A_{q_1,q_2} = 0.5 + 0.5 \cdot cor\left(E^{(q_1)}, E^{(q_2)}\right)$. The red area represents strong correlations between modules (absolute value is obtained representing either positive or negative correlation), and the green area represents modules are more distinct. Since they are not distinct, we decided to merge them (shown in Figure 21).

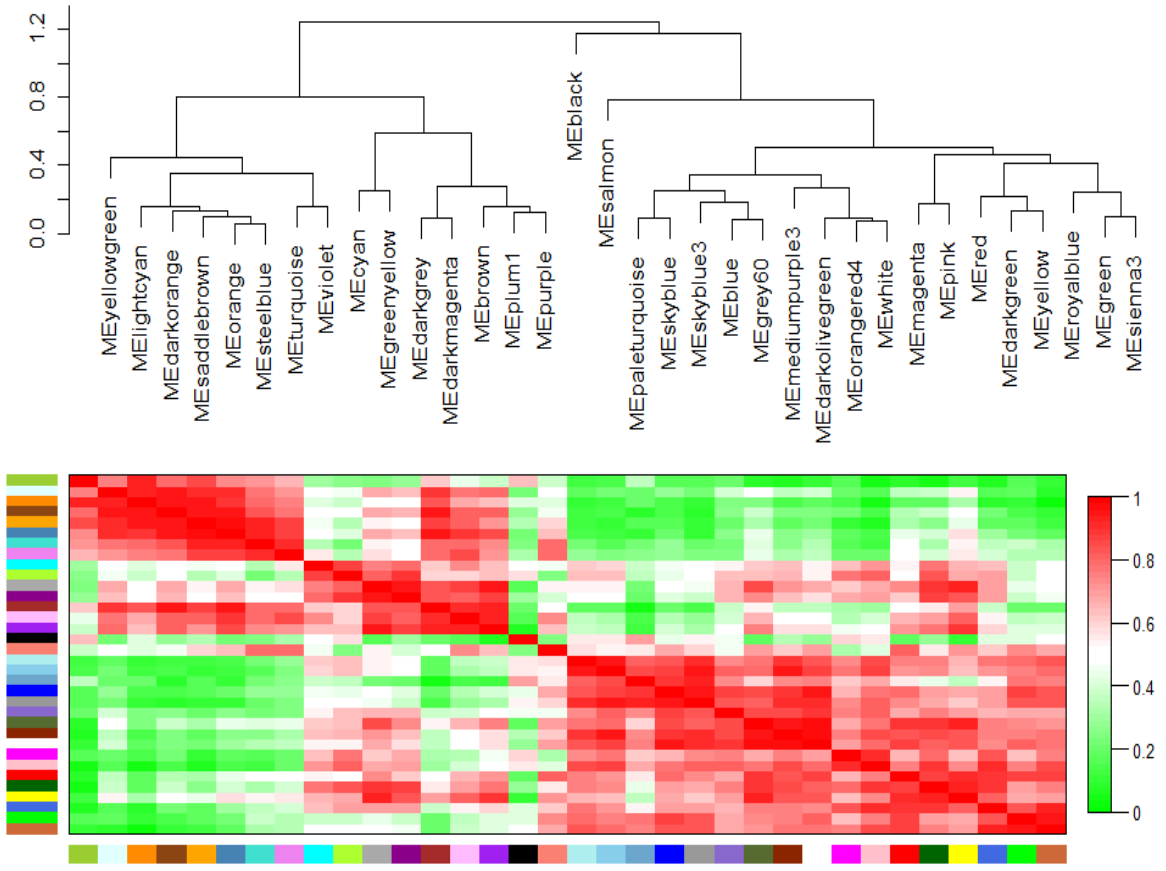


**Figure 21: Visualization of the eigengenes network representing the relationship among the modules.**

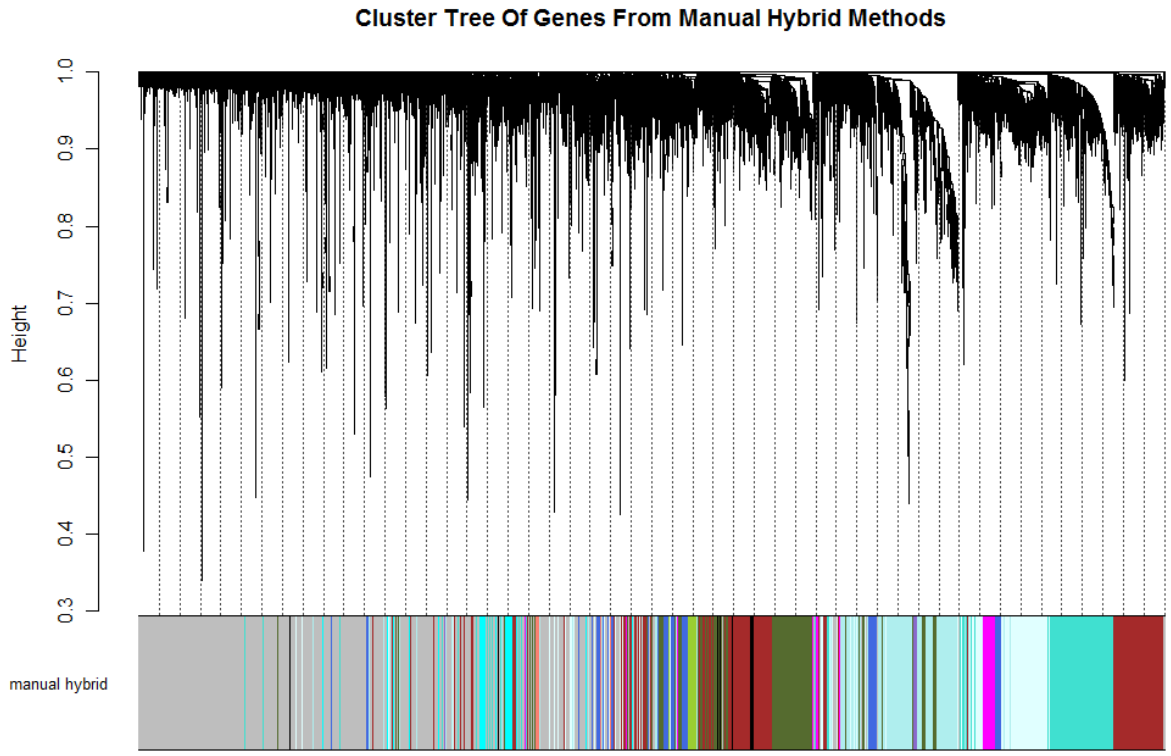


**Figure 22: Dendrogram after correlated modules are merged.**

We used step-by-step network construction and module detection method to produce a hierarchical clustering tree (dendrogram) of genes, and we grouped genes into 13 modules. In the dendrogram, each leaf, that is a short vertical line, corresponds to a gene. Branches of the dendrogram group together densely interconnected and highly co-expressed genes. Module identification amounts to the identification of individual branches. Modules are indicated by the color bands below the dendrogram. Table 29 shows the frequency of genes belongs to each module.

| Module | Frequency |
|---|---|
| Black | 204 |
| Brown | 2309 |
| Cyan | 421 |
| Dark Olivegreen | 1636 |
| Grey | 6439 |
| Light Cyan | 1685 |
| Magenta | 352 |
| Medium Purple | 101 |

| | |
|---|---|
| Pale Turquoise | 1776 |
| Royal Blue | 775 |
| Salmon | 103 |
| Turquoise | 1547 |
| Yellow Green | 181 |

**Table 29: Frequency of genes belongs to each module**

We use module eigengene to represent each module. The module eigengene of a given module is defined as the first principal component of the standardized expression profiles. Module eigengene is considered as the best summary of the standardized module expression data, and we used it to summarize the gene expression profiles of a given module [80-81].

We would like to see how many of these modules are related to disease status. Since we have 34 samples, in which 17 are carcinoma samples and 17 are normal samples. Among 13 modules, except modules Salmon and Grey, all the remaining 11 modules are significantly related to disease, which is illustrated by barplot of Figure 22. In Figure 23, the first color-band shows the result of step-by-step network construction and module detection. The second color-band visualizes the module significance from the disease status. “Red” color corresponds to positive gene significance (GS), green color indicates negative significance, and white color indicates no gene significance. Color saturation corresponds to GS strength. Turquoise module contains many genes that are highly negatively correlated with the disease status. Similarly, the third color band annotates genes significance by the disease status.

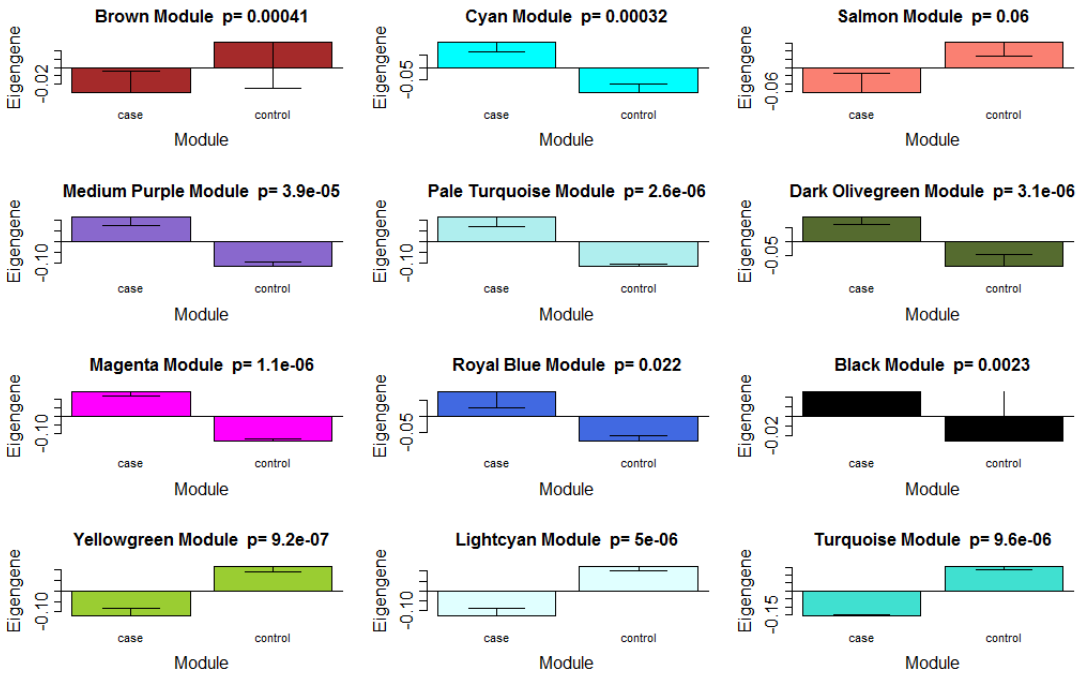


**Figure 23: Module Significance by Disease Status. 11 mRNA modules are significantly related to disease status. Only Salmon module and Grey module (not shown here) do not behave gene significance.**

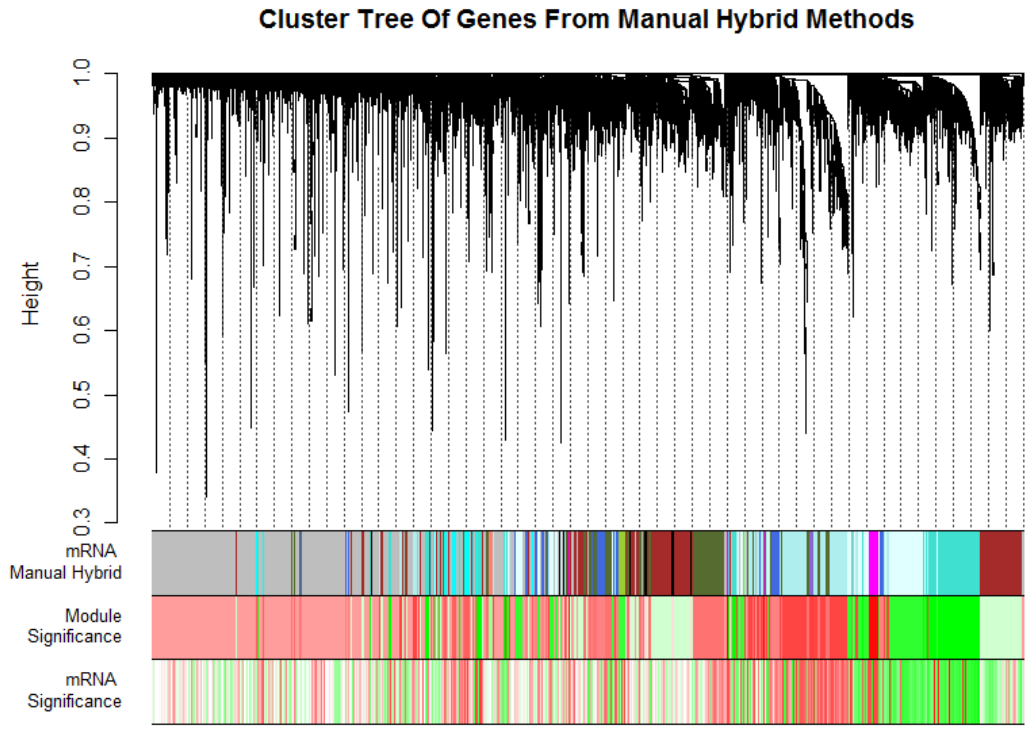


**Figure 24: Hierarchical cluster tree (average linkage, dissTOM) of 17529 genes.**

**(The colour bands provide a simple visual look of module assignment. The first colour-band shows the result of step-by-step network construction and module detection. The second colour-band visualizes the module significance from the disease status. The third colour-band visualizes the genes significance based on measurement from disease status. "red indicates a high positive correlation and "green" indicates a high negative correlation with disease status.)**

Thus, mRNAs were clustered based on the dissimilarity matrix, so that mRNAs with similar profiles were clustered together. mRNAs closer on the dendrogram are in a close distance and share similar patterns and are thus likely functionally related.

In total there are 319 miRNA in the renal carcinoma data set. Out of 319 miRNAs 257 could be lcoated to their putative genes targts through "TargetScan". For each miRNA, we run the Fisher exact test by a 2*2 contingency table, where row is if genes belong to a given module or not, and column is if these genes are putative targets or not. We evaluated whether the putative genes are differentially present between a given module and other modules. We are particularly interested in the 1$^{st}$ cell that belongs to a given module and putative gene targets. If the Fisher's exact test is significant, it suggests there is a different pattern between a given module and other modules in terms of putative gene targets. For this module, it has higher frequencies of mRNA targets than we expect by chance alone.

We compute the Pearson correlation between miRNA expression level and mRNA eigengenes from each module, and choose those miRNA/mRNA who are highly correlated (absolute correlation > = 0.70). These constraints remove spurious matches, reducing relatively speculative "putative" seed match based mRNA targets in "TargetScan" databases to a highly robust subset of direct functional targets. We retained only those miRNAs who are significant in the Fisher's exact test, and are also strongly correlated with module eigengenes. We found that 4 mRNAs are significant and their correlations with module eigengenes are shown in Table 19.

| miRNA | Strongly | Correlation | P-value | |
|---|---|---|---|---|

| | correlated to mRNA Module | between miRNA and Module Eigengene | from Fisher's Exact Test | Belong to miRNA Module |
|---|---|---|---|---|
| Hsa_miR_200c | Mediumpurple | -0.74 | 0.01 | brown |
| Hsa_miR_25 | Magenta | 0.73 | 0.01 | blue |
| Hsa_miR_362_5p | Magenta | -0.77 | 0.04 | brown |
| Hsa_miR_7 | Yellowgreen | -0.74 | 0.03 | blue |

**Table 30: 4 miRNAs who are significant in the Fisher exact test, and are also strongly correlated**

Figure 25 shows their relationship by the Fisher's exact test. The cell counts that belong to a given module and putative genes have been reported, so does the p-value from the Fisher's exact test. Figure 26 illustrates the heatmap of the correlation and p-values between 4 miRNAs and mRNA modules.


**Figure 25: Fisher's exact tests between gene modules and miRNA. The counts that belong to module i and putative targets and p-value have been reported.**

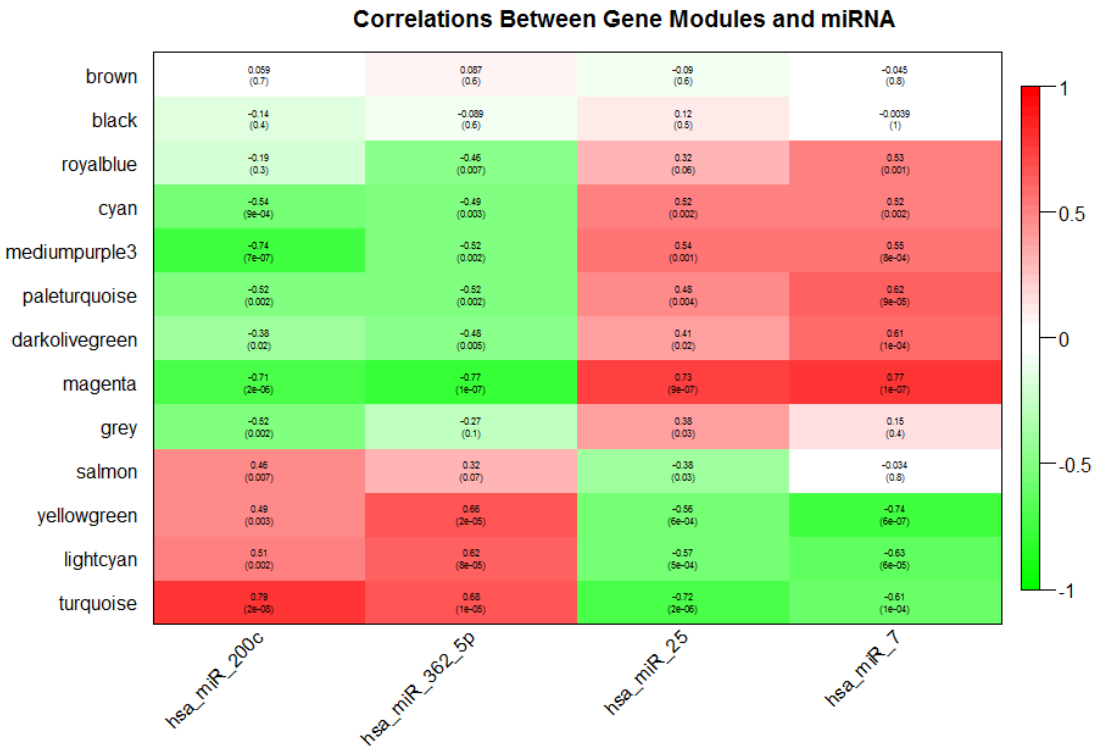


**Figure 26: Heatmap of correlations and p-values between miRNA and mRNA module eigengenes. Each cell indicates the correlation and p-values of miRNA and mRNA eigengenes.**

There are 319 miRNA in the renal carcinoma data set. The raw dataset is already log 2 transformed and quantiled normalized. Out of 319 miRNA 257 could be located by their putative genes through "targetscan". We applied signed co-expression step-by-step networks construction and module detection to analyze these 257 miRNAs across 34 expression arrays, and we define 6 modules. By default, we choose the power =12 for signed gene network analysis and we specify the minimum module size is 10.

Figure 27 shows the hierarchical clustering dendrogram of miRNAs and they are grouped into 6 modules. The first colour-band shows the result of step-by-step network construction and module detection. Turquoise and grey modules contain most genes in the module. The second colour-band visualizes module significance based on measurement from the disease status. The

third colour-band visualizes the genes significance based on measurement from disease status. Red indicates a high positive correlation and green indicates a high negative correlation with disease status. Table 31 shows the frequency of miRNAs in each module.

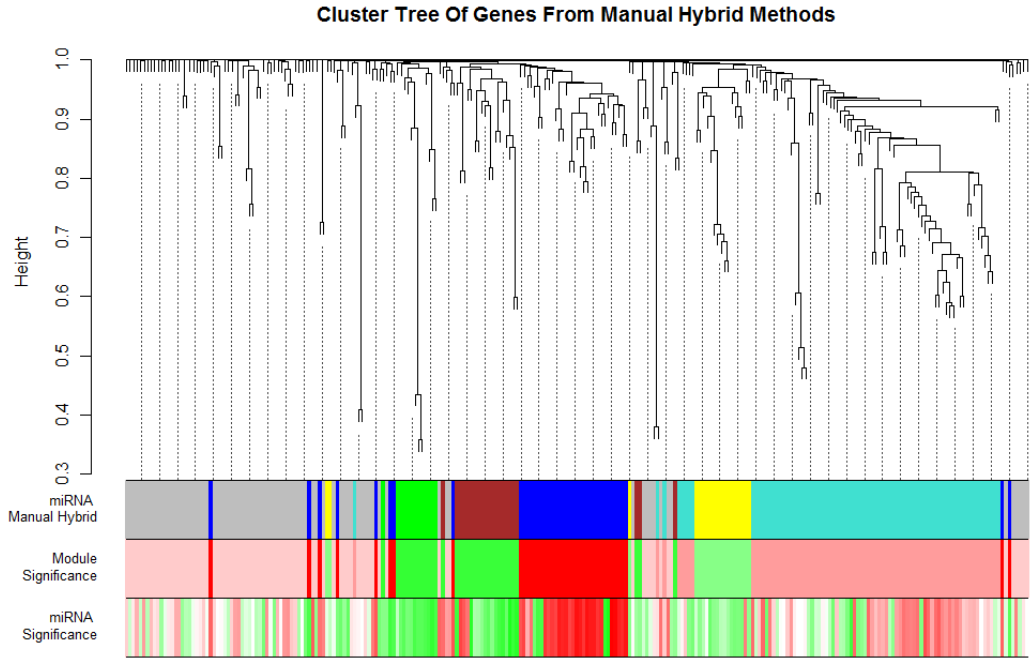


**Figure 27: Hierarchical cluster tree (average linkage, dissTOM) of 257 genes.**
**(The colour bands provide a simple visual look of module assignment. The first colour-band shows the result of step-by-step network construction and module detection. The second colour-band visualizes module significance based on measurement from disease status. The third colour-band visualizes the genes significance based on measurement from disease status. "red indicates a high positive correlation and "green" indicates a high negative correlation with disease status.)**

| Module | Frequency |
|---|---|
| Blue | 41 |
| Brown | 22 |
| Green | 13 |
| Grey | 83 |
| Turquoise | 79 |
| Yellow | 19 |

**Table 31: Frequency of miRNA in Each Module**

Figure 28 reveals the relationship between miRNA modules, mRNA modules and the disease status. The first row illustrates the miRNA significance with respect to the disease status. The first column shows the gene (mRNA) significance with regard to disease status. The remaining cells reveal the correlation and p-value between miRNA modules and gene modules.

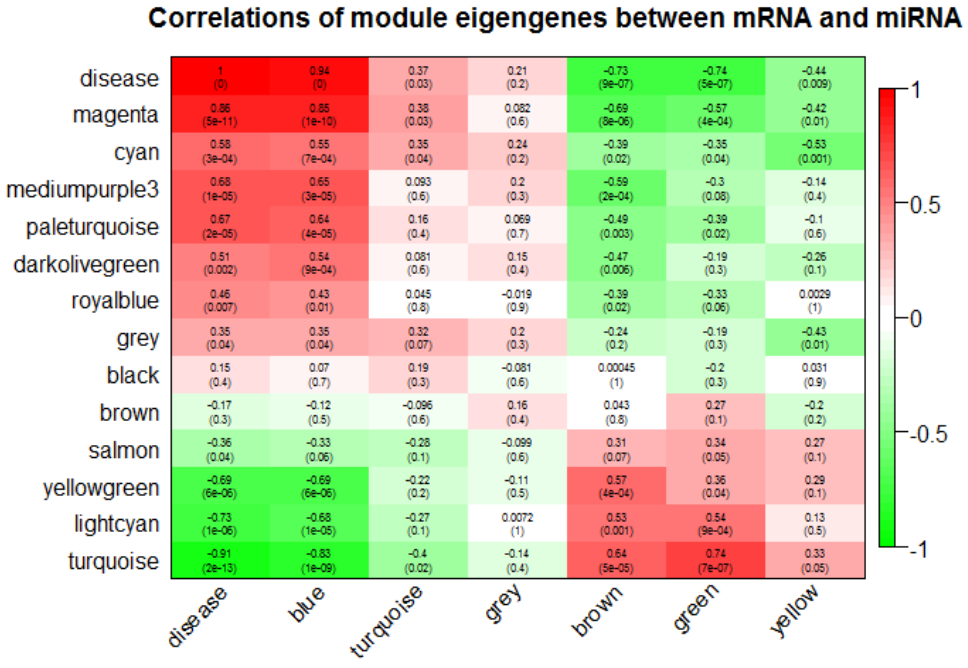


**Figure 28: Correlations of genes modules, miRNA modules, and disease status.**
**(The first row illustrates the miRNA significance with respect to the disease status. The first column shows the gene (mRNA) significance with regard to disease status. The remaining cells reveal the correlation and p-value between miRNA modules and gene modules.)**

We use least absolute shrinkage and selection operator (LASSO) regression to explore the relationship between miRNA and mRNA, so that the most informative miRNA with respect to our interest could be finalized. In the LASSO regression, the predictors are miRNA module eigengene and the outcome is the mRNA eigengene. We performed LASSO regression for each mRNA module respectively. The result is shown in table 32.

We find that a miRNA "hsa_miR_25" is significantly anti-correlated with magenta mRNA module. "hsa_miR_25" belongs to the blue miRNA module, which is also predictive to magenta mRNA module through the LASSO regression.

| Module | $R^2$ | Selected Modules | Parameter Estimates |
|---|---|---|---|
| Black | 0.35 | Blue | -0.75 |
| | | Turquoise | -0.05 |
| | | Brown | -0.58 |
| | | Green | 0.29 |
| | | Yellow | -0.73 |
| | | Grey | 0.31 |
| Brown | 0.47 | Turquoise | 0.09 |
| | | Brown | -0.38 |
| | | Yellow | -0.41 |
| | | Grey | 0.14 |
| Cyan | 0.33 | Blue | 1.07 |
| | | Turquiose | -0.57 |
| | | Brown | 0.98 |
| | | Green | 0.34 |
| | | Yellow | 0.29 |
| | | Grey | -0.31 |
| Dark Olivegreen | 0.57 | Blue | 1.10 |
| | | Turquoise | -0.10 |
| | | Green | 0.36 |
| | | Yellow | 0.27 |
| | | Grey | 0.03 |
| Grey | 0.49 | Blue | 0.70 |
| | | Yellow | 0.22 |
| Lightcyan | 0.42 | Blue | 0.88 |
| | | Turquoise | -0.19 |
| | | Green | 0.38 |
| | | Yellow | -0.03 |
| Magenta | 0.74 | Blue | 0.65 |
| | | Brown | -0.16 |
| Medium Purple | 0.28 | Blue | 0.54 |
| | | Yellow | 0.25 |
| | | Grey | -0.07 |
| Pale Turquoise | N/A | N/A (no modules are selected) | N/A |
| Royalblue | 0.53 | Blue | -0.70 |
| | | Brown | 0.17 |
| | | Green | -0.19 |

| | | | |
|---|---|---|---|
| Salmon | 0.60 | Blue | -0.80 |
| | | Turquoise | -0.09 |
| | | Green | 0.08 |
| | | Yellow | -0.37 |
| | | Grey | 0.14 |
| Turquoise | 0.79 | Blue | -0.43 |
| | | Turquoise | -0.21 |
| | | Brown | 0.25 |
| | | Green | 0.29 |
| | | Yellow | -0.16 |
| Yellowgreen | 0.20 | Brown | -0.09 |
| | | Yellow | -0.28 |

**Table 32: LASSO Regression where the outcome is module eigengene (mRNA), and the predictors are eigengene from 6 miRNA modules**

For each of the 13 mRNA modules, we have applied DAVID software to perform enrichment analysis and reported all the significant functions (p-value of Benjamini test < = 0.05) within each module. Table 33 shows the most significant term for each module.

| Module | Category | Term | Benjamini P-value |
|---|---|---|---|
| Black | SP_PIR_KEYWORDS | phosphoprotein | 2.70E-07 |
| Brown | SP_PIR_KEYWORDS | alternative splicing | 1.50E-12 |
| Cyan | SP_PIR_KEYWORDS | phosphoprotein | 3.90E-07 |
| Dark Olivegreen | SP_PIR_KEYWORDS | phosphoprotein | 9.60E-05 |
| Grey | SP_PIR_KEYWORDS | alternative splicing | 9.10E-35 |
| LightCyan | SP_PIR_KEYWORDS | phosphoprotein | 4.80E-09 |
| Magenta | GOTERM_BP_FAT | immune response | 8.90E-10 |

| | | | |
|---|---|---|---|
| Medium Purple3 | SMART | IG | 7.90E-03 |
| Pale | UP_SEQ_FEATURE | compositionally biased region:Pro-rich | 1.90E-12 |
| Royal Blue | SP_PIR_KEYWORDS | phosphoprotein | 2.60E-11 |
| Salmon | SP_PIR_KEYWORDS | leber hereditary optic neuropathy | 2.20E-08 |
| Turquoise | SP_PIR_KEYWORDS | ion transport | 1.10E-06 |
| Yellowgreen | SP_PIR_KEYWORDS | phosphoprotein | 1.30E-03 |

**Table 33: the most significant term in each mRNA module.**

## 5.6 Multiple Testing Problem

In microarray experiments a common issue is the identification of differentially expressed genes. The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of no association between the expression levels. A typical microarray experiment measures expression levels for thousands of genes simultaneously, large multiplicity problem are generated. When testing for potential differential expression across those conditions, each gene is considered independently from one another. i.e. a t-test is performed on each gene separately. A false positive or type I error, is defined as genes are falsely called differentially expressed when they are not. In this situation if type I error ($\alpha$) is 0.05, it suggests 5% probability that gene's mean expression level in one condition is different than the other by chance alone. If 10,000 genes are tested, 500 genes

would be called significant by chance alone. This is why it is important to correct the p-value of each gene when performing a statistical test on a group or genes.

A number of recent articles have addressed the question of multiple testing in microarray. Such as Dudoit et al[17], Efron et al[82], Golub et al[1], Kerr, Martin and Churchill[83], Manduchi et al[84], Tusher, Tibshirani and Chu[85], Westfall, and Young[86].

## 5.7 Multiple Testing Correction

Multiple testing correction adjusts the individual p-value for each gene to keep the overall error rate (or false positive rate) to be less than or equal to the user-specified p-value cutoff or error rate.

Consider the problem of testing simultaneously m null hypotheses $H_j$, j =1, …, m, and denote R the number of rejected hypotheses. In the frequentist setting, this could be summarized by the following table. The specific m hypotheses are assumed to be known in advance, the numbers $m_0$ and $m_1 = m - m_0$ of true and false null hypotheses are known parameter, R is an observable random variable and S, T, U and V are unobservable random variables. In the microarray context, there is a null hypothesis $H_j$ for each gene j and rejection of $H_j$ corresponds to declaring the gene j is differentially expressed. V is type I errors or false positives, and T is type II error or false negatives.

| Number of | Number not rejected | Number rejected | |
|---|---|---|---|
| True null hypotheses | U | V (type I error) | $m_0$ |
| Non-true null hypotheses | T (type II error) | S | $m_1$ |
| | m-R | R | m |

A few generalizations to the multiple testing situation are possible and particularly the following two concepts are important:

1) The family-wise error rate (FWER) is defined as the probability of at least type I error, that is $FWER = \Pr(V \geq 1)$

2) The false discovery rate (FDR) of Benjamini and Hochberg is the expected proportion of type I errors among the rejected hypotheses, that is $FDR = E(Q)$, by definition $Q = V/R$ if $R > 0$ and $0$ if $R = 0$

A multiple testing procedure is said to control a particular type I error rate at level α, if this error rate is less than or equal to α when the given procedure is applied to produce a list of R rejected hypotheses. FWER is controlled at level α by a particular multiple testing procedure if $FWER \leq \alpha$ and similarly for the other definitions of type I error rates.

In this paper, we will introduce three commonly used multiple testing corrections:

1) Bonferroni correction

The p-value of each gene is multiplied by the number of genes in the gene list. If the corrected p-value is still below the error rate, the gene will be significant:

Corrected p-value = p-value * n (number of genes in test) < 0.05.

As a consequence, if testing 1000 genes at a time, the highest accepted individual p-value is 0.00005, making the correction very stringent.

Bonferroni is a very conservative method that simply divides the type I error by the number of tests performed. But it does not take into account the dependence structure between genes.

2) Westfall and Young Permutation

Bonferron is a single- step procedure, where each p-value is corrected independently. The Westfall and Young[86] permutation method takes advantage of the dependence structure between genes, by permuting all the genes at the same time.

The Westfall and Young permutation follows a step-down procedure and combines with a bootstrapping method to compute the p-value distribution. The algorithm is as follows:

1. For the original data, order the observed test statistics for each gene such that

$$|t_{s1}| \geq |t_{s2}| \geq \cdots |t_{sm}|$$

| row | gene | $\|t_{si}\|$ |
|---|---|---|
| 1 | 3 | 7.1 |
| 2 | 2 | 3.4 |
| 3 | 5 | 2.8 |
| 4 | 4 | 0.2 |
| 5 | 1 | 0.1 |

2.For the $b^{th}$ permutation, b=1,…B, permute the n columns of the data matrix X, and compute test statistics $t_{i,b}$ for each hypothesis. Below is the original data where T stands for treatment and C is the control.

| gene | T | T | T | C | C | C |
|---|---|---|---|---|---|---|
| 1 | . | . | . | . | . | . |
| 2 | . | . | . | . | . | . |
| 3 | . | . | . | . | . | . |

Permutation divides dataset into new “treatment” and “control” group.

| gene | T | C | T | C | C | T |
|---|---|---|---|---|---|---|
| 1 | . | . | . | . | . | . |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | . | . | . | . | . | . |
| 3 | . | . | . | . | . | . |

3.Compute $u_{i,b} = \max_{l=1,..m} | t_{sl,b} | = \max( | t_{si,b} |, | t_{si+1,b} |, \cdots | t_{sm,b} | )$, by assigning

$$u_{m,b} = | t_{sm,b} |$$
$$u_{i,b} = \max( u_{i+1,b,} | t_{si,b} | )$$

| row | gene | $\|t_{si}\|$ | $\|t_{si,b}\|$ | $u_{i,b}$ |
|---|---|---|---|---|
| 1 | 3 | 7.1 | 1.8 | $u_{1,b}= \max(u_{2,b},\|t_{s1,b}\|)=3.0$ |
| 2 | 2 | 3.4 | 2.1 | $u_{2,b}= \max(u_{3,b},\|t_{s2,b}\|)=3.0$ |
| 3 | 5 | 2.8 | 3.0 | $u_{3,b}= \max(u_{4,b},\|t_{s3,b}\|)=3.0$ |
| 4 | 4 | 0.2 | 0.8 | $u_{4,b}= \max(u_{5,b},\|t_{s4,b}\|)=1.3$ |
| 5 | 1 | 0.1 | 1.3 | $u_{5,b}=\|t_{s5,b}\|=1.3$ |

4.Find the indicator function I ($u_{i,b} \geq |t_{si}|$), repeated the above step B times and the adjusted p-values are: $\tilde{p}^*_{si} = \frac{\sum_{b=1}^{B} I(u_{i,b} \geq | t_{si} |)}{B}$

| row | gene | $\|t_{si}\|$ | $\|t_{si,b}\|$ | $u_{i,b}$ | I $(u_{i,b} \geq \|t_{si}\|)$ | $\sum$ | $p_{si}^* = \sum/1000$ |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 7.1 | 1.8 | 3.0 | 0 | 48 | 0.048 |
| 2 | 2 | 3.4 | 2.1 | 3.0 | 0 | 145 | 0.145 |
| 3 | 5 | 2.8 | 3.0 | 3.0 | 1 | 138 | 0.138 |
| 4 | 4 | 0.2 | 0.8 | 1.3 | 1 | 876 | 0.876 |
| 5 | 1 | 0.1 | 1.3 | 1.3 | 1 | 935 | 0.935 |

$$\tilde{p}^*_{s1} = \tilde{p}^*_{s1}$$
$$\tilde{p}^*_{si} = \max( \tilde{p}^*_{si}, \tilde{p}^*_{si-1}) \qquad for\ 2 \leq i \leq m$$

5. Setting

| row | gene | $\|t_{si}\|$ | $p_{si}^{*}$ = ∑/1000 | $P_{si}^{*}$ |
|---|---|---|---|---|
| 1 | 3 | 7.1 | 0.048 | 0.048 |
| 2 | 2 | 3.4 | 0.145 | 0.145 |
| 3 | 5 | 2.8 | 0.138 | 0.145 |
| 4 | 4 | 0.2 | 0.876 | 0.876 |
| 5 | 1 | 0.1 | 0.935 | 0.935 |

This is the final adjustment of p-values, to keep the ranks consistent with the original p-values.

3) Benjamini and Hochberg False Discovery Rate

A different approach to multiple testing was proposed in 1995 by Benjamini and Hochberg. They proposed a less conservative approach which calls for controlling the expected proportion of type I errors among the rejected hypotheses – the false discovery rate, FDR.

Specifically, FDR is defined as $FDR = E(Q)$, where $Q = V/R$ if $R > 0$ and $0\ if\ R = 0$. i.e. $FDR = E(V/R \mid R > 0) \Pr(R > 0)$.

Benjamini and Hochberg[87] derived the following step-up procedure to strong control the FDR for independent test statistics.

1.The observed unadjusted p-values are ranked from smallest to the largest $P_1 \le P_2 \cdots \le P_m$

| Test | Raw |
|---|---|
| 1 | 0.0001 |
| 2 | 0.0058 |
| 3 | 0.0132 |
| 4 | 0.0289 |

| | |
|---|---|
| 5 | 0.0498 |
| 6 | 0.0911 |
| 7 | 0.2012 |
| 8 | 0.5718 |
| 9 | 0.8912 |
| 10 | 0.9011 |

2.Find k=max {k: p(k) ≤ kα/m}. In this example, the largest p-value to satisfy p(k) ≤ kα/m is p(3)

| **Test** | **Raw** | **P(k) ≤ kα/m=k*0.05/10** |
|---|---|---|
| 1 | 0.0001 | 0.005 |
| 2 | 0.0058 | 0.01 |
| 3 | 0.0132 | 0.015 |
| 4 | 0.0289 | 0.02 |
| 5 | 0.0498 | 0.025 |
| 6 | 0.0911 | 0.03 |
| 7 | 0.2012 | 0.035 |
| 8 | 0.5718 | 0.04 |
| 9 | 0.8912 | 0.045 |
| 10 | 0.9011 | 0.05 |

3.Thus, reject $H_0$ corresponding to p(1), p(2), and p(3)

4.Adjusted p-values are: $\widetilde{p}_{(j)} = \min_{k=j,\cdots m}\{\min(\frac{m}{k} p_{(k)},1)\}$

i.e.

$$\widetilde{p}_{(m)} = p_{(m)}$$

$$\widetilde{p}_{(m-1)} = \min\{\widetilde{p}_{(m)}, \frac{m}{m-1} p_{(m-1)}\}$$

$$\widetilde{p}_{(m-2)} = \min\{\widetilde{p}_{(m-1)}, \frac{m}{m-2} p_{(m-2)}\}$$

$$\vdots \quad \vdots \quad \vdots$$

$$\widetilde{p}_{(1)} = \min\{\widetilde{p}_{(2)}, mp_{(1)}\}$$

$$\tilde{p}_{(10)} = p_{(10)} = 0.9011$$

$$\tilde{p}_{(9)} = \min(\tilde{p}_{(10)}, \frac{10}{9} p_9) = \min(0.9011, 0.9902) = 0.9011$$

$$\tilde{p}_{(8)} = \min(\tilde{p}_{(9)}, \frac{10}{8} p_8) = \min(0.9011, 0.7148) = 0.7148$$

……

Example:

| Test | Raw | False Discovery Rate |
|---|---|---|
| 1 | 0.0001 | 0.0010 |
| 2 | 0.0058 | 0.0290 |
| 3 | 0.0132 | 0.0440 |
| 4 | 0.0289 | 0.0723 |
| 5 | 0.0498 | 0.0996 |
| 6 | 0.0911 | 0.1518 |
| 7 | 0.2012 | 0.2874 |
| 8 | 0.5718 | 0.7148 |
| 9 | 0.8912 | 0.9011 |
| 10 | 0.9011 | 0.9011 |

In the microarray setting, where thousands of tests are performed simultaneously and a fairly large number of genes are expected to be differentially expressed, FDR controlling procedures present a promising alternative to FWER approaches.

## 5.8 WGCNA Alleviates Multiple Testing Problems

In our research, we first applied WGCNA to construct network and detected modules for highly correlated genes and we found 13 modules. We then identified the strong correlation

between miRNA module eigengenes and mRNA module eigengenes. Through "TargetScan", we located the putative gene targets for each miRNA. We next evaluate for each miRNA, whether its putative gene targets are differentially present between a given module and other modules by using the Fisher's exact test. We retained miRNAs who are significant in the Fisher's exact test, and are strongly correlated with mRNA module eigengenes.

Our method greatly alleviates the multiple testing problems that plague standard gene-centric methods [88]. Instead of testing the relationship between thousands of genes and individual miRNA, it focuses on the relationship between a few modules (here 13) and individual miRNA. Because the modules may correspond to biological pathways, focusing the analysis on module eigengenes (and equivalently intra-modular hub genes) amounts to a biologically motivated data reduction scheme. WGCNA starts from the level of thousands of genes, identifies clinically interesting gene modules, and finally uses intra-modular connectivity to suggest suitable targets. Because the expression profiles of intra-modular hub genes inside an interesting module are highly correlated typically dozens of targets. Although these targets are statistically equivalent, they may differ in terms of biological plausibility or clinical utility. In many applications, the list of module hub genes may be further narrowed down based on (*i*) biological plausibility based on external gene (ontology) information which is explored here in our paper. (*ii*) the availability of protein biomarkers for further validation.

We have carried out a simulation plan to compare our WGCNA method with the traditional methods, i.e. those methods that directly relate gene expression with miRNA expression to look for their correlation. We found our method provides much smaller false discovery rate.

## 5.9 Simulation

For a fixed cut-off value d for a test statistic $z_j$, we can obtain the true or realized FDR and its estimates as[74],

$$FDR(d) = \pi_0 FP(d) / T\hat{P}(d)$$
$$FD\hat{R}(d) = \hat{\pi}_0 F\hat{P}(d) / T\hat{P}(d)$$

Where $\pi_0$ is the proportion of equally expressed genes among all genes, and $\hat{\pi}_0$ is its estimator. FP is the number of false positive genes, i.e., the number of equally expressed genes but claimed as differentially expressed genes, $F\hat{P}$ is the estimated number of false positive genes. $T\hat{P}(d)$ is the total number of genes claimed as differentially expressed by a certain criteria.

In order to obtain $F\hat{P}$, we need to estimate the distribution of the test statistic $Z_i$ under the null hypothesis that gene *i* is an equally expressed gene. Rather than assuming a parametric distribution for the null distribution of $Z_i$, a class of non-parametric methods has been proposed to estimate it empirically. The idea is to impute the data and calculate the null statistics $Z_i$ in the same way as calculating $Z_i$, but based on the permuted data.

Under the null hypothesis, the empirical distribution of the null statistics can be used to approximate the null distribution. In our context, the dimension of gene matrix X is 34* 17529. Rows correspond to samples where the first 17 are controls and the remaining 17 are cancers. Columns represent for each gene.

Under null hypothesis, we can permute the gene data by randomly permuting the order of each column. And we performed a large number of random permutations (B times). Calculating the same test statistic from the b-*th* permuted data results in the null statistic $Z_i^{(b)}$ for b = 1, …, B and i = 1, …, G. For any given cut off value d > 0, if we claim any gene i satisfying $|Z_i| > d$ to be significant, we estimate the true positive (TP) numbers and false positive (FP) numbers as:

$$\hat{TP}(d) = \#\{i : |Z_i| > d\}$$

$$\hat{FP}(d) = \sum \#\{i : |Z_i|^{(b)} > d\} / B$$

We plug $\hat{TP}(d)$ and $\hat{FP}(d)$ to calculate FDR(d) and $\hat{FDR}(d)$. However, owing to the difficulty of assigning p-values, the estimation of $\pi_0$ remains challenging. Guo and Pan[89] pointed out that permutation method would over-estimate p-values, and thus lead to the overestimation of $\pi_0$. Dalmasso et al[90] use a method to estimate only an upper bound of $\pi_0$. Since estimation of $\pi_0$ is itself an unsettled research question, and it's not our focus in this context, we bypass it in the simulation. For the simulation result, we use true $\pi_0$ which represents the ideal performance of the standard method to replace $\hat{\pi}_0$. We regard this is a constant and all the simulation results are proportional to it.

Simulation 1: WGCNA method

The true datasets: a miRNA dataset that includes 257 miRNA, and a mRNA dataset that includes 17529 genes. We applied step-by-step network construction and module detection to the gene matrix and 13 modules are defined. We choose the miRNAs who are strongly correlated with each module eigengene (the absolute correlation > =0.7). By "TargetScan" we locate the putative genes for each miRNA. For each miRNA, we evaluated whether its putative gene targets are differentially present between a given module and other modules by the Fisher's exact test. We retained miRNAs who are significant in the Fisher's exact test, and are strongly correlated with module eigengenes. We observed 4 miRNAs falling into our criteria.

Then we permute column of gene matrix for one time and apply the same procedure to the permuted gene matrix. We count the number of significant miRNAs. Following this, we repeat this procedure for 1000 times. We sum the number of significant miRNAs through the

permutations and divide them by 1000. In total 36 miRNAs have been chosen. Thus we obtain $\hat{FP}(d) = 36/1000 = 0.036$. The false discover rate = 0.036/4 = 0.009.

When we change the cut off value of absolute correlation to be greater or equal to 0.6, we observe 21 miRNAs that are strongly correlated with their putative genes targets. Through 1000 permutation, we find in total 303 miRNAs are significant. So the estimated false positive rate is $\hat{FP}(d) = 303/1000 = 0.303$. The false discover rate is calculated as: 0.303/21 = 0.014. Similarly, when we change the cut off value of absolute correlation to be greater or equal to 0.5, we observe 47 miRNAs that are strongly correlated with their putative genes targets. Through 1000 permutation, we find in total 1427 miRNAs are significant. The estimated false positive rate is $\hat{FP}(d) = 1427/1000 = 1.427$ and the false discover rate is calculated as: 1.427/47 = 0.031.

| Fisher's Exact Test | Absolute Correlation Between miRNA Expression Level and mRNA Module Eigengenes | False Discovery Rate |
|---|---|---|
| P-value <= 0.05 | >=0.7 | 0.009 |
| P-value <= 0.05 | >=0.6 | 0.014 |
| P-value <= 0.05 | >=0.5 | 0.031 |

**Table 34: Simulated FDR from our method**

Simulation 2: traditional method

Following the tradition method such as MMIA[40], we choose the miRNAs who are strongly correlated with gene expression (the absolute correlation > = 0.7). We then go through "TargetScan" to locate the putative genes for each miRNA. We observed 37 miRNAs that are strongly correlated with their putative genes targets.

Then we permute column of gene matrix for one time and apply the same procedure to the permuted gene matrix. We count the number of miRNAs that are strongly correlated with genes.

Following this, we repeat the same procedure for 1000 times. We sum the number of significant miRNAs through the permutations and divide them by 1000. In total 2541 miRNAs have been chosen. Thus we obtain $\hat{FP}(d) = 2541/1000 = 2.541$. The false discover rate is calculated as: 2.541/37 = 0.068.

When we change the cut off value of absolute correlation to be greater or equal to 0.6, we observed 88 miRNAs that are strongly correlated with their putative genes targets. Through 1000 permutation, we find in total 16912 miRNAs are significant. So the estimated false positive rate is $\hat{FP}(d) = 16912/1000 = 16.912$. The false discover rate is calculated as: 16.912/88 = 0.192. Similarly, when we change the cut off value of absolute correlation to be greater or equal to 0.5, we observed 164 miRNAs that are strongly correlated with their putative genes targets. Through 1000 permutation, we find in total 82515 miRNAs are significant. The estimated false positive rate is $\hat{FP}(d) = 82515/1000 = 82.515$ and the false discover rate is calculated as: 82.515/164 = 0.503.

| Absolute Correlation Between miRNA and mRNA Expression | False Discovery Rate |
|---|---|
| >=0.7 | 0.068 |
| >=0.6 | 0.192 |
| >=0.5 | 0.503 |

**Table 35: Simulated FDR from traditional method**

Table 36 below lists and compares overall false discovery rates produced from both our method and the traditional method. Our method provides much smaller false discovery rates than the tradition method. By applying WGCNA to define modules for highly correlated genes first then relate module eigengenes to miRNAs, we greatly relieve the multiple testing problems and lower the false discovery rates inherent in microarray data analysis.

| Absolute Correlation Between miRNA and Module Eigengene/Gene Expression | False Discovery Rate from Our Method | False Discovery Rate from Traditional Method |
|---|---|---|
| >=0.7 | 0.009 | 0.068 |
| >=0.6 | 0.014 | 0.192 |
| >=0.5 | 0.031 | 0.503 |

**Table 36: Comparison of False Discovery Rate from Both Methods**

## 5.10 Results and Conclusion

Weighted gene co-expression network analysis (WGCNA) is a systems biology method that describes the correlation patterns among genes across microarray samples. We used it to find clusters (modules) of highly correlated genes, to summarize clusters using the module eigengene, and to relate modules to disease status by using eigengene network methodology.

We first perform the step-by-step network construction and module detection of highly correlated genes. We found 13 modules. We then identify the strong correlation between miRNA and module eigengenes. Through "TargetScan", we locate the putative genes for each miRNA. We then evaluate for each miRNA, whether its putative genes are differentially present between a given module and other modules by using the Fisher's exact test. We retained miRNAs who are significant in the Fisher exact test, and are strongly correlated with module eigengenes.

Next we relate modules to disease status by using eigengene network methodology, and we find that 11 modules are strongly related with disease status. Within these modules, enrichment analyses are implemented by DAVID.

We also run step-by-step network construction and module detection of miRNAs and define 6 modules. We use LASSO regression to explore the relationship between miRNA and

mRNAs. The predictors are miRNA module eigengene and the outcome is the mRNA module eigengene. We find that a miRNA "hsa_miR_25" is significantly anti-correlated with magenta mRNA module. "hsa_miR_25" belongs to the blue miRNA module, which is also predictive to magenta mRNA module through the LASSO regression. Its putative gene targets are found and integrated from the renal carcinoma dataset.

The advantage of using the step-by-step network construction and module detection by the WGCNA package is: WGCNA package partitioned tons of genes into a few modules (clusters) in which genes are closely related and share similar features and are thus likely functionally related. It does not require the number of clusters to be pre-specified and has nice visualization properties with dendorgram and heatmap. By first identifying the modules from WGCNA, it saves the huge work of computing the correlation between each miRNA and genes.

In addition, our method greatly alleviates the multiple testing problems that plague standard gene-centric methods. Instead of testing the relationship between thousands of genes and individual miRNA, it focuses on the relationship between a few modules (here 13) and individual miRNA. Because the modules may correspond to biological pathways, focusing the analysis on module eigengenes (and equivalently intramodular hub genes) amounts to a biologically motivated data reduction scheme. WGCNA starts from the level of thousands of genes, identifies clinically interesting gene modules, and finally uses intramodular connectivity to suggest suitable targets.

We carried out the simulation plan and compared the false discovery rate produced from our method with other currently available software such as MMIA. Our method successfully relieves the multiple comparison problem and provides much smaller false discovery rate.

Weighted gene co-expression network analysis and gene ontology data provides a novel integrative view of miRNA and their prediction targets. We described a simple and reliable method to identify direct putative genes of miRNA in a renal carcinoma dataset. But this method could be extended to other cancer datasets. We applied the WGCNA package and applied constraints that miRNA and mRNA are strongly correlated, to remove spurious matches and identify a subset of putative gene targets for each miRNA, and greatly reduce the false discovery rate commonly existing in microarray setting. This gives a guidance to better understand the relationship between miRNAs and genes as well as their joint behaviors.

## References


1. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
2. Dalma-Weiszhausz, D.D., et al., *The affymetrix GeneChip platform: an overview.* Methods Enzymol, 2006. **410**: p. 3-28.
3. *Multiplex Analysis with Bead-Based Assays. Multimetrix GmbH.*
4. Amararunga, D., *Exploration and analysis of DNA microarray and protein array data.* John Wiley and Sons, New Jersey., 2004.
5. Wu, W., et al., *Evaluation of normalization methods for cDNA microarray data by k-NN classification.* BMC Bioinformatics, 2005. **6**: p. 191.
6. Yang, H., et al., *A segmental nearest neighbor normalization and gene identification method gives superior results for DNA-array analysis.* Proc Natl Acad Sci U S A, 2003. **100**(3): p. 1122-7.
7. Smyth, G.K. and T. Speed, *Normalization of cDNA microarray data.* Methods, 2003. **31**(4): p. 265-73.
8. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. **32 Suppl**: p. 496-501.
9. Bilban, M., et al., *Normalizing DNA microarray data.* Curr Issues Mol Biol, 2002. **4**(2): p. 57-64.
10. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.
11. Li, C. and W. Hung Wong, *Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.* Genome Biol, 2001. **2**(8): p. RESEARCH0032.
12. Chen, Y.J., et al., *Normalization methods for analysis of microarray gene-expression data.* J Biopharm Stat, 2003. **13**(1): p. 57-74.
13. Edwards, D., *Non-linear normalization and background correction in one-channel cDNA microarray studies.* Bioinformatics, 2003. **19**(7): p. 825-33.
14. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
15. Workman, C., et al., *A new non-linear normalization method for reducing variability in DNA microarray experiments.* Genome Biol, 2002. **3**(9): p. research0048.
16. Fang, Y., et al., *A model-based analysis of microarray experimental error and normalisation.* Nucleic Acids Res, 2003. **31**(16): p. e96.
17. Dudoit, S., et al., *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Statistica Sinica, 2002. **12**(1): p. 111-139.
18. Kroll, T.C. and S. Wolfl, *Ranking: a closer look on globalisation methods for normalisation of gene expression arrays.* Nucleic Acids Res, 2002. **30**(11): p. e50.
19. Tseng, G.C., et al., *Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.* Nucleic Acids Res, 2001. **29**(12): p. 2549-57.
20. Cui, X., M.K. Kerr, and G.A. Churchill, *Transformations for cDNA microarray data.* Stat Appl Genet Mol Biol, 2003. **2**: p. Article4.

21. Zien, A., et al., *Centralization: a new method for the normalization of gene expression data.* Bioinformatics, 2001. **17 Suppl 1**: p. S323-31.
22. Kepler, T.B., L. Crosby, and K.T. Morgan, *Normalization and analysis of DNA microarray data by self-consistency and local regression.* Genome Biol, 2002. **3**(7): p. RESEARCH0037.
23. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.
24. Clark, D. and L. Russell, *Molecular Biology Made Simple and Fun.* Vienna, IL: Cache River Press, 1997.
25. Gonick, L. and M. Wheelis, *A cartoon Guide to Genetics.* New York: Harper Collins. , 1991.
26. Yang, Y.H., et al., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Res, 2002. **30**(4): p. e15.
27. Wang, Y., et al., *Iterative normalization of cDNA microarray data.* IEEE Trans Inf Technol Biomed, 2002. **6**(1): p. 29-37.
28. Cleveland, W.S., *Robust locally weighted regression and smoothing scatterplots.* Journal of the American Statistical Association, 1979. **74**: p. 829-836.
29. Cleveland, W.S. and S.J. Devlin, *Locally Weighted Regression: An approach to regression analysis by local fitting.* . Journal of the American Statistical Association, 1988. **83**(403): p. 596-610.
30. Venables, W.N. and B.D. Ripley, *applied statistics with S-PLUS.* Springer, 3rd edition, 1999.
31. Yang, Y.H., *Normalization for cDNA microarray data, preprint #589* Statistics Dept, UC Berkeley, Jan 2001, 2002.
32. Jemal, A., et al., *Cancer statistics, 2008.* CA Cancer J Clin, 2008. **58**(2): p. 71-96.
33. Youlden, D.R., S.M. Cramb, and P.D. Baade, *The International Epidemiology of Lung Cancer: geographical distribution and secular trends.* J Thorac Oncol, 2008. **3**(8): p. 819-31.
34. Doll, R. and R. Peto, *The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today.* J Natl Cancer Inst, 1981. **66**(6): p. 1191-308.
35. Walser, T., et al., *Smoking and lung cancer: the role of inflammation.* Proc Am Thorac Soc, 2008. **5**(8): p. 811-5.
36. Nix, B. and D. Wild, *Calibration curve-fitting. The immunoassay handbook, 2nd edition.* Nature publishing group, New York, NY., 2001.
37. Lin, L.I., *A concordance correlation coefficient to evaluate reproducibility.* Biometrics, 1989. **45**(1): p. 255-68.
38. Park, T., et al., *Evaluation of normalization methods for microarray data.* BMC Bioinformatics, 2003. **4**: p. 33.
39. Qin, L.X., *An integrative analysis of microRNA and mRNA expression--a case study.* Cancer Inform, 2008. **6**: p. 369-79.
40. Nam, S., et al., *MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W356-62.
41. Reinhart, B.J., et al., *The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.* Nature, 2000. **403**(6772): p. 901-6.

42. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
43. Lim, L.P., et al., *The microRNAs of Caenorhabditis elegans.* Genes Dev, 2003. **17**(8): p. 991-1008.
44. Lai, E.C., et al., *Computational identification of Drosophila microRNA genes.* Genome Biol, 2003. **4**(7): p. R42.
45. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.
46. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
47. Ambros, V., *The functions of animal microRNAs.* Nature, 2004. **431**(7006): p. 350-5.
48. Volinia, S., et al., *A microRNA expression signature of human solid tumors defines cancer gene targets.* Proc Natl Acad Sci U S A, 2006. **103**(7): p. 2257-61.
49. Calin, G.A. and C.M. Croce, *MicroRNA signatures in human cancers.* Nat Rev Cancer, 2006. **6**(11): p. 857-66.
50. Farh, K.K., et al., *The widespread impact of mammalian MicroRNAs on mRNA repression and evolution.* Science, 2005. **310**(5755): p. 1817-21.
51. Lewis, B.e.a., *Conserved seed pairing, often flanked by adenosine, indicates that thousands of human Genes are microRNA targets.* . Cell, 2005. **120**: p. 15-20.
52. Xiao, F., et al., *miRecords: an integrated resource for microRNA-target interactions.* Nucleic Acids Res, 2009. **37**(Database issue): p. D105-10.
53. Sethupathy, P., B. Corda, and A.G. Hatzigeorgiou, *TarBase: A comprehensive database of experimentally supported animal microRNA targets.* RNA, 2006. **12**(2): p. 192-7.
54. Nam, S., et al., *miRGator: an integrated system for functional annotation of microRNAs.* Nucleic Acids Res, 2008. **36**(Database issue): p. D159-64.
55. Maragkakis, M., et al., *DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W145-8.
56. Sales, G., et al., *MAGIA, a web-based tool for miRNA and Genes Integrated Analysis.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W352-9.
57. Rathmell, W.K. and S. Chen, *VHL inactivation in renal cell carcinoma: implications for diagnosis, prognosis and treatment.* Expert Rev Anticancer Ther, 2008. **8**(1): p. 63-73.
58. Calzada, M.J. and L. del Peso, *Hypoxia-inducible factors and cancer.* Clin Transl Oncol, 2007. **9**(5): p. 278-89.
59. Ficarra, V., et al., *Original and reviewed nuclear grading according to the Fuhrman system: a multivariate analysis of 388 patients with conventional renal cell carcinoma.* Cancer, 2005. **103**(1): p. 68-75.
60. Gottardo, F., et al., *Micro-RNA profiling in kidney and bladder cancers.* Urol Oncol, 2007. **25**(5): p. 387-92.
61. Jung, M., et al., *MicroRNA profiling of clear cell renal cell cancer identifies a robust signature to define renal malignancy.* J Cell Mol Med, 2009. **13**(9B): p. 3918-28.
62. Juan, D., et al., *Identification of a microRNA panel for clear-cell kidney cancer.* Urology, 2010. **75**(4): p. 835-41.

63. Boer, J.M., et al., *Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array.* Genome Res, 2001. **11**(11): p. 1861-70.
64. Petillo, D., et al., *MicroRNA profiling of human kidney cancer subtypes.* Int J Oncol, 2009. **35**(1): p. 109-14.
65. Liu, H., et al., *Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell Renal Cell Carcinoma.* BMC Syst Biol, 2010. **4**: p. 51.
66. Smalheiser, N.R. and V.I. Torvik, *A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions.* BMC Bioinformatics, 2004. **5**: p. 139.
67. Sethupathy, P., M. Megraw, and A.G. Hatzigeorgiou, *A guide through present computational approaches for the identification of mammalian microRNA targets.* Nat Methods, 2006. **3**(11): p. 881-6.
68. Rajewsky, N., *microRNA target predictions in animals.* Nat Genet, 2006. **38 Suppl**: p. S8-13.
69. Megraw, M., et al., *miRGen: a database for the study of animal microRNA genomic organization and function.* Nucleic Acids Res, 2007. **35**(Database issue): p. D149-55.
70. Krek, A., et al., *Combinatorial microRNA target predictions.* Nat Genet, 2005. **37**(5): p. 495-500.
71. Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.
72. Enright, A.J., et al., *MicroRNA targets in Drosophila.* Genome Biol, 2003. **5**(1): p. R1.
73. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
74. Tibshirani, e.a., *Cluster validation by prediction strength.* Journal of Computational and Graphical Statistics, 2005. **14**: p. 511-28.
75. Kerr, M.K. and G.A. Churchill, *Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments.* Proc Natl Acad Sci U S A, 2001. **98**(16): p. 8961-5.
76. Dudoit, S. and J. Fridlyand, *A prediction-based resampling method for estimating the number of clusters in a dataset.* Genome Biol, 2002. **3**(7): p. RESEARCH0036.
77. Horvath, S., *Weighted Network Analysis: Applications in Genomics and Systems Biology.* Springer, 1st edition, 2011.
78. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.
79. Kaufman, L. and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* . New York: John Wiley and Sons, Inc. , 1990.
80. Langfelder, P. and S. Horvath, *Eigengene networks for studying the relationships between co-expression modules.* BMC Syst Biol, 2007. **1**: p. 54.
81. Horvath, S. and J. Dong, *Geometric interpretation of gene coexpression network analysis.* PLoS Comput Biol, 2008. **4**(8): p. e1000117.
82. Efron, B., et al., *Microarrays and their use in a comparative experiment.* . Technical Report 2000-37B/213, Dept. Statistics, Stanford Univ. , 2000.
83. Kerr, M.K., M. Martin, and G.A. Churchill, *Analysis of variance for gene expression microarray data.* J Comput Biol, 2000. **7**(6): p. 819-37.

84. Manduchi, E., et al., *Generation of patterns from gene expression data by assigning confidence to differentially expressed genes.* Bioinformatics, 2000. **16**(8): p. 685-98.
85. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
86. Westfall, P.H., D.V. Zaykin, and S.S. Young, *Multiple tests for genetic effects in association studies.* Methods Mol Biol, 2002. **184**: p. 143-68.
87. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.* J. R. Statist. Soc. , 1995. **57**(1): p. 289-300.
88. Horvath, S., et al., *Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17402-7.
89. Guo, X. and W. Pan, *Using weighted permutation scores to detect differential gene expression with microarray data.* J Bioinform Comput Biol, 2005. **3**(4): p. 989-1006.
90. Dalmasso, C., P. Broet, and T. Moreau, *A simple procedure for estimating the false discovery rate.* Bioinformatics, 2005. **21**(5): p. 660-8.