INVESTIGATION OF THE EFFECTS OF MAS5, RMA AND GCRMA PREPROCESSING METHODS ON AN AFFYMETRIX ZEBRAFISH GENECHIP® DATASET USING STATISTICAL AND NETWORK PARAMETERS

A THESIS SUBMITTED TO THE DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS AND THE INSTITUTE OF ENGINEERING AND SCIENCE OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

By

AHMET RAŞİT ÖZTÜRK

JANUARY 2010

TO MY MELİKE & MY FAMILY

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. Özlen Konu

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Assoc. Prof. Dr. Işık G. Yuluğ

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. Tolga Can

Approved for the Institute of Engineering and Science

_____

Prof. Dr. Mehmet Baray
Director of Institute of Engineering and Science

ABSTRACT

INVESTIGATION OF THE EFFECTS OF MAS5, RMA AND GCRMA PREPROCESSING METHODS ON AN AFFYMETRIX ZEBRAFISH GENECHIP® DATASET USING STATISTICAL AND NETWORK PARAMETERS

Ahmet Raşit Öztürk

MSc, in Molecular Biology and Genetics

Supervisor: Assist. Prof. Özlen Konu

January, 2010, 125 pages

Microarray data preprocessing is an important determinant of the accuracy and repeatability of expression profiling studies. Recent studies have focused on comparison of preprocessing methodologies using differential expression analysis of spike-in datasets and qRT-PCR confirmations. Other approaches include comparison of array-wise and probe-wise correlation and of selected gene network parameters. However, zebrafish GeneChip datasets have not been used in such comparisons; furthermore, detailed analysis of upregulated and downregulated gene sets with respect to known network parameters are not well characterized across different preprocessing methodologies. In this study we re-analyzed a public zebrafish hypoxia microarray dataset (GSE4989; Marques et al. 2008) using MAS5, RMA, and gcRMA methods. Comparisons were made in terms of differentially expressed gene sets and defined network parameters, namely, clustering coefficient, degree distribution, and betwenness centrality. Our findings indicated that gcRMA and RMA exhibited greater similarity to each other in terms of differentially expressed genes, and network parameters. In addition, the network analysis demonstrated that upregulated and downregulated gene sets had distinct network structures; downregulated probesets had greater clustering coefficients and degree distributions for positively correlated probesets in all three preprocessing methods. However, gcRMA and RMA methods accentuated this difference further than MAS5 did, suggesting that preprocessing methods differ in their modulation of gene expression network structure. A selected

group of probesets that showed invariant network structure parameters across RMA, gcRMA and MAS5 was determined and analyzed functionally for the zebrafish hypoxia dataset. The results of this thesis suggest that preprocessing methods may alter network structure of the datasets differentially with respect to upregulated and downregulated gene sets. Accordingly, it might be beneficial to filter differentially expressed genes that are robust to such network topology modulation to increase the repeatability of gene sets.

ÖZET

MAS5, RMA VE GCRMA ÖN İŞLEME METOTLARININ BİR AFFYMETRIX ZEBRAFISH GENECHIP® VERİ KÜMESİ ÜZERİNE ETKİLERİNİN İSTATİSTİKSEL PARAMETRELER VE AĞ PARAMETRELERİ KULLANILARAK ARAŞTIRILMASI

Ahmet Raşit Öztürk

Moleküler Biyoloji ve Genetik Yüksek Lisansı

Tez Yöneticisi: Yrd. Doç. Dr. Özlen Konu

Ocak, 2010, 125 sayfa

Mikrodizi veri ön işlemesi, ifade profili çıkarma çalışmalarının kesinlik ve tekrar edilebilirliğinin önemli bir belirleyici faktörüdür. Güncel çalışmalar, farklılaşmış ifade analizleri kullanılarak ön işleme metodolojilerinin kontrol probları içeren mikrodizi veri kümeleri ve qRT-PCR doğrulamaları yoluyla karşılaştırılması üzerine yoğunlaşmıştır. Diğer yaklaşımlarsa dizi ve prob boyunca karşılaştırmalarla birlikte, seçilmiş gen ağ parametrelerinin karşılaştırmalarını içermektedir. Ancak zebrabalığı GeneChip veri kümelerinde henüz böyle bir karşılaştırma kullanılmamıştır, ayrıca, bilinen ağ parametreleriyle ilgili olarak anlatımı artan veya azalan gen öbeklerinin detaylı analizi farklı ön işleme metodolojileri üzerinden iyi bir biçimde tanımlanmamıştır. Bu çalışmada bir zebrabalığı hipoksi mikrodizi veri seti (GSE4989; Marques et al. 2008) MAS5, RMA ve gcRMA metodları ile analiz edilmiştir. Karşılaştırmalar, farklı ifade edilen gen öbekleri açısından "öbeklenme katsayısı", "derece dağılımı" ve "aradalık merkeziyeti" olarak adlandırılan ağ parametreleri referans alınarak yapılmıştır. Bulgularımız gcRMA ve RMA metotlarının farklı ifade edilen genler ve ağ parametreleri açısından daha yüksek bir benzerlik gösterdiğini işaret etmektedir. Bunun yanı sıra ağ analizi, anlatımı artan ve azalan gen öbeklerinin farklı ağ yapılarına sahip olduğunu, pozitif korelasyon gösteren probsetler açısından anlatımı azalan probsetlerin her üç ön işleme metodunda da daha yüksek öbeklenme katsayısı değerleri ve ağ grafiğinde daha fazla bağlantıya sahip olduğunu göstermiştir. Bu durum MAS5 metoduna göre gcRMA ve RMA metotları tarafından işlenen verilerde daha ön plana çıkmıştır. Bu durum da ön işleme metotlarının gen ifade

ağlarının yapılarını şekillendirmekte farklı etkilerinin olduğunu göstermektedir. RMA, gcRMA ve MAS5 ile ön işlemeye tabi tutulan verilerden oluşturulan ağlar arasında ağ topolojisi bakımından en az değişiklik gösteren bir probset öbeği seçilmiş ve zebrabalığı hipoksi veriseti için fonksiyonel olarak analiz edilmiştir. Bu tezin sonuçları, ağ yapılarının anlatımı artan ve azalan gen kümeleri açısından ön işleme metotları tarafından değiştirildiğini önermektedir. Buna göre, farklı ifade edilen gen kümelerinin tespitinin tekrarlanabilirliğini arttırmak için ağ topolojilerindeki değişimlere dayanıklı genleri filtrelemek yararlı olabilir.

ACKNOWLEDGEMENTS

I would like to thank to Assist. Prof. Özlen Konu for her supervision, valuable suggestions and for sharing her experiences on bioinformatics during my undergraduate and master studies. It has always been a privilege to work with such a special person.

I would like to express my thanks to Assoc. Prof. Işık G. Yuluğ and Assist. Prof. Tolga Can for their helpful comments for my thesis study.

I am indebted to especially Muammer Üçal, Ceren Sucularlı, Onur Kaya and Rümeysa Bıyık on behalf of all my friends for providing a stimulating environment in the lab. I would like to thank them on their friendship and support.

I would like to attend my thanks to Fatih Semerci for his support for my thesis and my thoughts.

Many thanks to MBG family who were with me during my studies.

Lastly, I would like to thank my family and my better half, Melike, for everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

DEG	:	Differentially Expressed Gene
gcRMA	:	GC Robust Multi-chip Average
GUI	:	Graphical User Interface
MAS5	:	Microarray Suite version 5.0
MM	:	Mismatch
PM	:	Perfect Match
qRT-PCR	:	Quantitative Real Time Polymerase Chain Reaction
RMA	:	Robust Multi-chip Average

CHAPTER 1: INTRODUCTION

1.1. Microarrays in human disease

Every year millions of people die because of various cancer types and other diseases. The reason for most treatment to fail is the lack of understanding of the nature of the disease. Since the cellular mechanisms causing a disease is complex, high-throughput methods are necessary in order to show the relationships between different genes or proteins (Draghici 2003). Bio-molecular interaction networks, which can be partly assessed through characterization of gene expression profiles and can be relatively easily obtained using microarray techniques, provide extensive information to understand the organization and interactivity of a biological system involved in disease (Draghici 2003).

The main advantage of microarrays is the ability of quickly and inexpensively getting the gene expression profile of a certain tissue or cell type at the transcriptome level (Seidel and Niessner 2008). Today, nearly all of the known transcriptome can be mapped and expression values can be obtained (Yauk *et al.* 2007). This notion gains importance especially in the case of cancer and drug treatment where a certain condition affects the behavior (or in other terms, expression pattern) of a group of cells. Transcriptome studies might be beneficial in the areas of drug development, diagnosis, comparative genomics, and functional genomics (Thorgeirsson *et al.* 2006; Wiltgen *et al.* 2007). It's also important to increase understanding of mechanisms of physiological and cellular processes that might contribute to disease process, such as hypoxia, ischemia, and oxidative stress. Micorarray studies allow for expression profiling at a large scale; and functional annotation of the gene lists obtained from such analyses leads to novel signaling pathway characterizations involved in different pathologies (Quackenbush 2002).

1.2. Zebrafish as a model for human disease and gene networks

Model organisms are widely used in biomedical research (Amatruda *et al.* 2008; Gonzalez-Nunez *et al.* 2009). Zebrafish is a model organism that help the researcher find the effect of a condition within a short time due to its high reproductive capacity

and relatively short embryonic and larval development (Dooley *et al.* 2000). In addition, zebrafish has been shown to share common properties in terms of molecular response in human cancer models such as liver cancer (Berghmans *et al.* 2005). These characteristics make zebrafish an excellent good organism for studying human disease.

A limited number of protein-protein interaction and gene regulatory networks of zebrafish have been generated in the literature and some web tools generated for the visualization and the basic analysis of those networks also include zebrafish datasets. In a recent study (Sorathiya *et al.* 2009), a microarray dataset has been normalized using the RMA method and a gene expression network was generated. Network topology analysis revealed a group of genes which may have a role in the early stages of vasculogenesis in zebrafish. In another study (Webb *et al.* 2009), visualization of an *in silico* network was performed to identify the effects of amphetamine on gene expression. Other two studies (Bacha *et al.* 2009; Jupiter *et al.* 2009) focus on the development of web-based tools for network visualization and analysis using zebrafish protein-protein interaction and gene expression networks for demonstration of the tools.

1.3. DNA microarrays

Several types of microarrays are available for a wide variety of purposes (e.g., DNA, protein, and tissue microarrays). DNA microarrays can be separated into two types as cDNA and oligo arrays where a cDNA or a 25-80 base long oligonucleotide is spotted microscopically onto a solid surface (Schena *et al.* 1995). Another way of producing microarrays is through lithography where oligonucleotides are synthesized on the array using a special technique of light masking (Pease *et al.* 1994).

The working principle of microarrays is the hybridization of two oligonucleotides which is detected by the fluorescence emitted during the hybridization process (Draghici 2003). The signal from the oligonucleotide pair is read from special optic devices where the intensity of the signal represents the amount of hybridization for a specific oligonucleotide, namely the amount of gene expression. However, the intensity value can only be interpreted with respect to control spots or intensity values of a set of housekeeping genes (Millenaar *et al.* 2006; Irizarry, Bolstad, *et al.* 2003).

Getting a relative gene expression value dictates the necessity of utilizing a normalization method in order to make the comparison between different arrays possible (Geller *et al.* 2003). Normalization is also necessary to handle the differences between arrays in terms of RNA extraction, labeling, hybridization, and scanning as well as other types of systematic error (Draghici 2003).

1.4. Preprocessing of Microarrays

Although microarray technology promises great advantages, it also has its problems (Quackenbush 2002; Hill *et al.* 2001; Bilban *et al.* 2002; Boes *et al.* 2005; Kreil *et al.* 2005; Grewal *et al.* 2007; Canales *et al.* 2006; Steinhoff *et al.* 2006; Zakharkin *et al.* 2005; Smyth *et al.* 2003). One of the handicaps is the fact that one can obtain relative gene expression values, not the actual ones using the hybridization based techniques (Schena *et al.* 1995; Geller *et al.* 2003). Second, since the experimental procedures from RNA isolation to hybridization are error prone and also affected by the experience of the technician, non-biological noise is inserted into the microarray data (Bolstad *et al.* 2003). In addition, the distribution of different probe sets on different microarray platforms makes the comparison of different studies more difficult (Irizarry, Hobbs, *et al.* 2003). Although the first problem mentioned depends on the nature of microarrays, the latter ones can be solved through data transformations (Geller *et al.* 2003). The data preprocessing is a necessary step to remove the non-biological noise from the real signal as much as possible. Several preprocessing and normalization techniques have been proposed including RMA (Irizarry, Hobbs, *et al.* 2003), gcRMA (Wu *et al.* 2004), MAS5.0 (Hubbell *et al.* 2002), dChip (Li *et al.* 2001), PLIER (Affymetrix), and others (Shedden *et al.* 2005).

Despite the fact that proposed preprocessing techniques are useful for reducing systematic technical noise, there is no golden standard. One of the suggested methods is MAS5.0, a method proposed and suggested for Affymetrix for data preprocessing. In this method, mismatch probes, with a single mismatch from the perfect match probes, are taken into account for calculating the amount of true hybridization; and a scaling approach is used for data normalization (Hubbell *et al.* 2002). In addition, many studies use RMA since it has been shown to outperform other preprocessing techniques. RMA depends only on the perfect match probes from the microarray raw

data (Irizarry, Hobbs, *et al.* 2003; Katz *et al.* 2006; Bolstad *et al.* 2003; Chiogna *et al.* 2009). Another method called gcRMA has been introduced to improve the performance of RMA method by considering the effect of probe GC content (Wu *et al.* 2004). Although these methods are successfully utilized for the detection of differentially expressed genes (Draghici 2003; Shedden *et al.* 2005), it is also shown that the array platform, tissue type, sample size of the study, or numerous other conditions can affect the performance of the applied method (Giles *et al.* 2003).

1.5. Comparative analysis of Preprocessing Methods

1.5.1 Description of Methods

Preprocessing of microarray raw data is a three-step process for Affymetrix data that aims to result in the summed normalized signal intensity measurements (Draghici 2003; Irizarry, Hobbs, *et al.* 2003). Due to non-specific and false binding, filtering background noise from the data is the first crucial step. This step is called background correction. After filtering out the systematic noise from the data, normalization is applied. Normalization enhances the comparison of different data from different microarray experiments adjusting and scaling the main characteristics of the data, such as mean/median, distribution and/or standard deviation. After the normalization of the signal intensities of each nucleotide, the last step is summarization of the normalized values. Typically, a transcript is represented by 11 to 20 different short oligonucleotides and combining these multiple signal intensities is a crucial operation (Affymetrix). Summarization is usually the last step where signal intensities of multiple oligonucleotides, which represent a single transcript, are collected and summed into a single signal intensity value. Although some methods might have a different order or extra steps during preprocessing (Schuster *et al.* 2007), preprocessing methods that we refer in this thesis follow the steps mentioned above. A summary of each preprocessing method's approaches for background correction, normalization, and summarization is shown below (Lim *et al.* 2007) in Table 1. According to a study in 2009, MAS5 is the most preferred method in the literature for the preprocessing of Affymetrix HG-U133 array whereas RMA is the second preferred method (Kadota *et al.* 2009).

Table 1. Summary of each step of different preprocessing methods (modified from Lim *et al.*, 2007).

Method	Background correction	Normalization	Summarization
MAS5	MM Subtraction	Constant	Tukey biweight
RMA	RMA transformation	Quantile	Median polish
gcRMA	gcRMA transformation	Quantile	Median polish

1.5.2 MAS5

As shown in Table 1, MAS5 utilize PM-MM subtraction for background correction. For each oligonucleotide on the array, Affymetrix has designed a corresponding mismatch oligonucleotide in order to take the effect of non-specific binding into account. In addition, the method gives detection calls that represent the presence or absence of the expression of a gene. Using this property of arrays, MAS5 corrects the perfect match signal intensities using mismatch signal intensities for each oligonucleotide. MAS5 assumes a linear approximation of background correction. For the normalization step, this method uses constant scaling to normalize different arrays. Lastly, for summarization, Tukey's biweight approach is preferred (Hubbell *et al.* 2002). This summarization method is an efficient method for removing large median absolute deviations from the data. MAS5 removes background noise for each array independent from other arrays in the dataset. Thus, it is a single-chip method for preprocessing. So, preprocessing is not affected by addition or subtraction of arrays to the dataset (Binder *et al.* 2010).

Robust averages of PM-MM values are calculated in MAS5 method for background correction. However, variation of probes with low signals is increased. Also the subtraction adds extra noise to the data (Pepper *et al.* 2007). Obtaining MM values larger than PM values is another possible problem of MAS5 method, generally handled by using idealized MM values.

1.5.3. RMA

Due to the ineffective utilization of mismatch probes on the array, a new method was proposed depending just on the perfect match signal intensities (Irizarry, Hobbs, *et al.*

2003). Another property of this method is the utility of quantile normalization, which is a linear method for array-wise adjustment. Decomposition of the frequency distribution of signal intensities into an exponential signal and a Gaussian background distribution is the main approach of this method. Arrays are then normalized using quantile normalization, which scales the data across arrays in quantiles. Lastly, median polish, a summarization method, is used for getting a single signal intensity value for a transcript from multiple oligonucleotides (Lim *et al.* 2007). Median polish minimizes the residual log error. As a result, different signal intensities are transformed into one average distribution. RMA method decreases the variance of probes with low signal values (Binder *et al.* 2010).

1.5.4. gcRMA

gcRMA is the enhanced version of RMA method that uses GC content information of each nucleotide to calculate binding efficiency and thus, signal intensity. Since the strength of G-C hybridization is stronger than A-T, the GC content of an oligonucleotide affects the binding tendency of each oligonucleotide pair after washing the arrays (Wu *et al.* 2004). Normalization and summarization steps are the same as the RMA method. However, for background correction, gcRMA background correction method is applied (Lim *et al.* 2007). It is a multi array approach that is affected by addition or subtraction of arrays into a dataset. As a result, weighted averages of arrays are calculated and replaced with the original values (Binder *et al.* 2010).

1.5.5. Comparative preprocessing method studies and gene networks

In the literature, advantages and disadvantages of each preprocessing method is widely discussed (Kadota *et al.* 2008; Kadota *et al.* 2009; Hua *et al.* 2008; Verhaak *et al.* 2006; Qiu *et al.* 2005; Beyene *et al.* 2007; Liu *et al.* 2006; Reverter *et al.* 2005; Fujita *et al.* 2006; Harr *et al.* 2006; Shedden *et al.* 2005; Bolstad *et al.* 2003; Autio *et al.* 2009). Although a single method is not superior to others, it is concluded that the efficiency of the method is affected by the nature of the study (Verhaak *et al.* 2006). In addition, it is stated that MAS5 has more reliable results than other methods when applying a correlation-based statistical analysis like clustering analysis (Lim *et al.*

2007). In another case, RMA has been found superior when a list of significantly expressed genes was identified (Zakharkin *et al.* 2005). Although clustering and significantly expressed genes are investigated through different preprocessing methods, comparative gene network analysis has not been widely studied in the literature in this respect, to our knowledge. Most notably, Lim *et al*. (2007) has introduced the concept of normalization method comparisons in terms of reverse engineering gene networks. Another important landmark was a study performed by Ahn et al. (2009), which compared different networks with each other using specified network characteristics.

Although the generation of the network is usually performed using a correlation method like Pearson correlation (Selga *et al.* 2009; Baralla *et al.* 2009), network structure is thought to be robust and main parameters like clustering coefficient is a characteristic property of a network independent of the data (Strogatz 2001; Watts *et al.* 1998). Since it is shown that different tissue and cell types affect the efficiency of preprocessing method (Shedden *et al.* 2005; Gyorffy *et al.* 2009), it should be assessed if different preprocessing methods have significant effects on the network structure and the characteristics of network properties.

1.6. Zebrafish Microarrays

The microarray that is studied in this thesis belongs to the Affymetrix GeneChip Zebrafish Genome Array. The Affymetrix GeneChip Zebrafish Genome Array platform consists of 15,509 different probe sets for the detection of more than 14,900 zebrafish transcripts. Array is designed using the sequences from 2003 builds of RefSeq, Genbank, dbEST, and Unigene sequence databases. Each probeset includes 16 different 25-mer oligonucleotides long probes. Detection sensitivity scale is 1:100,000 (Affymetrix).

According to GEO database, there are currently 46 GeneChip Zebrafish Genome Array series representing 610 samples at the time this thesis is written. To our knowledge, there is no study in the literature focusing on the effects of normalization methods on Genechip Zebrafish Genome Array data. It is shown in the literature that different cell or tissue types can affect the statistical properties of data that affects the performance of preprocessing methods (Gyorffy *et al.* 2009; Shedden *et al.* 2005).

CHAPTER 2: AIMS AND STRATEGY

Microarray data analysis includes steps and methods such as preprocessing of raw data, differentially expressed gene identification, clustering, and visualization of gene regulatory networks. In this thesis, the aim is to investigate whether there is an advantage applying either the MAS5, RMA or gcRMA preprocessing algorithms to raw microarray data, using an exemplary dataset from zebrafish, to perform a more accurate gene profiling analysis.

Motivation behind the thesis is that although these three methods have been previously compared using human microarrays from different tissues and experiments, the effect of normalization on Affymetrix Zebrafish arrays has not been studied, to our knowledge. Since preprocessing is likely to affect the analysis results at different levels, better knowledge of the characteristic changes introduced by normalization and development of novel statistical analysis approaches would lead to a) a more suitable selection of normalization methods; and b) a more reliable interpretations of the obtained results.

To achieve the goal of this thesis, zebrafish Affymetrix microarray data on hypoxia (Marques *et al.* 2008) were obtained from GEO database of NCBI . The dataset was normalized using MAS5, RMA and gcRMA. Different analysis approaches were applied to compare the characteristic effects of the mentioned normalization/preprocessing methods: 1) assessment of data distribution; 2) assessment of differentially expressed genes; 3) determination and comparison of gene regulatory network characteristics. The reason for using a wide variety of analysis approaches is that different aspects of data characteristics could be influenced by each method. Thus, one can obtain a more thorough understanding of all facets of the problems and/or advantages associated with each method. For example, results of a t-test are affected by the extent of variation for each gene whereas the gene regulatory networks are affected by the relative rank of a gene among samples (Lim *et al.* 2007).

This thesis also aims to test whether upregulated and downregulated probesets have distinct network structure and properties (i.e., clustering coefficient, degree distribution, and betweenness centrality) in zebrafish hypoxia and normoxia

experiments; and whether these differences are affected differentially by application of microarray preprocessing algorithms. To test this hypothesis, positively correlated edges were used for network comparisons that were made between the up- and down-regulated probesets for each method and between methods using visualization techniques such as scatterplots and correlation analyses, and tests between real and randomized networks.

Finally, we also aimed to identify a subset of significantly differentially expressed probesets, with invariant network properties that are independent of the normalization method using the zebrafish hypoxia dataset. This study will help increase our understanding of the degree to which preprocessing can alter differentially expressed gene lists as well as the network structure in microarray datasets.

CHAPTER 3: MATERIALS AND METHODS

3.1. Dataset

In this study, a public microarray dataset (GSE4989; Marques *et al.* 2008) was downloaded from the NCBI GEO database (Barrett *et al.* 2005) to test the effects of normalization on differentially expressed gene lists and gene network characteristics. In GSE4989 expression series, gene expression in response to chronic constant hypoxia in the heart of adult zebrafish has been profiled. Data consisted of 10 Affymetrix GeneChip Zebrafish microarrays, five of which belonged to normoxia and the other five arrays were of the hypoxia group. Maques et al. (2008) identified 376 differentially expressed genes with a p value of 0.05 or less and a minimum fold change of 2. Although not mentioned in the paper, the preprocessing method used by the Marques et al. (2008) was RMA.

Hypoxia is the lack of getting enough oxygen into the body. It can be either generalized or tissue hypoxia depending on the severity and the location of the oxygen deprivation (Semenza 2001). Although mammals are not tolerant to hypoxia, it is known that some teleosts –such as zebrafish- have the ability to cope with extreme oxygen deprivation (Stecyk *et al.* 2004). The aim of the paper of Marques et al. (2008) has been to identify differentially expressed genes in hypoxia in order to understand the mechanism of developing such a tolerance. They have suggested that understanding the mechanism of zebrafish's response to chronic constant hypoxia in heart might have clinical implications in the future (Marques *et al.* 2008). In the present thesis, we aim to test whether such an expression profile is robust to differences in preprocessing methods particularly in terms of differentially expressed gene lists and gene network parameters.

3.2. Preprocessing Methods

MAS5, RMA, and gcRMA methods were utilized to preprocess the raw data obtained in the form of .CEL files. To automate the normalization process, an R script was written using the functions of Bioconductor R packages (APPENDIX A). Once .CEL files were read into a variable, data were preprocessed with the abovementioned

methods and written into a tab separated file for further use (APPENDIX A). In addition, each generated file was also saved in an MS Excel 97-2003 compatible format for further use. Since the rma() and gcrma() methods give log2 transformed data, log2 transformation also was applied to the data obtained from the mas5() function to make the preprocessed data values comparable (BioconductoR).

3.3. Assessment of Differentially Expressed Genes upon preprocessing

Bioinformatics Toolbox of MatLab 2008a was used for the current and the following steps. Once the preprocessed data have been read into variables, samples from each of the two different groups (e.g., hypoxia vs normoxia; GSE4989) were labeled for a t-test (APPENDIX B). mattest() function was used for calculation of p-values to identify differentially expressed genes by applying this function to each preprocessed dataset; and the probeset lists were obtained. Upon retrieval of the differentially expressed gene lists from each preprocessed data, union or intersection of these lists were obtained using a code that made use of the 'union' and 'intersect' functions (APPENDIX B).

3.4. Generation and comparison of gene expression networks

To generate gene expression networks, the Pearson correlation values (Equation 1) were calculated for each probe set pair within an experiment (e.g., GSE4989). Covariance of two datasets X and Y is divided by the multiplication of the standard deviations of X and Y.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Equation 1)}$$

Only the positively correlated gene pairs having an r value greater than or equal to 0.6 were considered to have an edge between them. The positively correlated gene pairs were analyzed in this study to simplify the network generation and decrease complexity. In this approach, each node represented a gene (probe set) and each edge represented coexpression of the two nodes. Resulting graph, a gene expression network, could be shown as an n-by-n matrix, where n is the number of nodes in the graph. A sparse matrix also was generated to be used for further investigation of network topology (APPENDIX C).

The distribution of correlation pairs across different correlation thresholds (Lim *et al*., 2007) were calculated and plotted using the plotting functions of MatLab. The average value of the positive correlation values from all pairwise probeset correlations (i.e., a mean edge correlation value) was compared between any two preprocessing method. Random networks (Lim *et al*., 2007) were generated to establish the extent of differences between observed and random correlations for each preprocessed dataset (APPENDIX C).

Since a differentially expressed gene list obtained from each preprocessed data contained different probe sets, the probesets of the union and those of the intersection lists were used to generate networks where all networks had the same number of nodes (APPENDIX D). Frequency distribution plots of bins of correlated pairs were compared among datasets obtained from different preprocessing methods for union and intersection datasets.

In this study, the following network measures were investigated: degree distribution (Newman 2003), clustering coefficient (Watts *et al.* 1998), and betweenness centrality (Freeman 1977).

Degree distribution is the number of edges in and out of a node.

$$C_i = \frac{\lambda_G(v)}{\tau_G(v)} \qquad \text{(Equation 2)}$$

C_i (clustering coefficient of the node *i*) is the proportion of neighboring subgraphs having 3 edges and 3 nodes to the number of neighboring subgraphs with 2 edges and 3 nodes (Equation 2; Watts *et al.* 1998).

$$C_B(v) = \sum_{\substack{s \neq v \neq t \in V \\ s \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad \text{(Equation 3)}$$

$C_B(v)$ value is the sum of all possible ratios of shortest paths passing the node *v* over the total number of those shortest paths (Equation 3; Freeman 1977) where σst(*v*) is the number of shortest paths from s to t crossing the node *v*. σst is the total number of shortest paths from s to t.

For these calculations, a set of related functions from the GAIMC BGL Toolbox 4.0 was utilized in Matlab (Anon 2008). This toolbox has been previously used in a brain network analysis study to calculate network properties (Sporns *et al.* 2007).

Comparison of networks from different preprocessing methods as well as comparison of a network with a randomized counterpart we used a series of methodologies. To find the conserved correlations (i.e., the conserved edges) among the networks of differently preprocessed data, union and intersection methods were used (APPENDIX D). Scatterplots were generated to visualize network parameters in a pairwise fashion between any two preprocessing method. Spearman correlation coefficients for a compared network parameter across probesets (Ahn *et al*., 2009) were calculated between preprocessing methods for the intersection and union datasets (APPENDIX G).

Paired t-tests were used to test the differences between any two preprocessing method for the three network parameters, separately. Furthermore, data from each preprocessing method was also compared with those from a randomized dataset using paired t-tests (APPENDIX B).

Upregulated and downregulated probeset lists were generated and plotted against the fold change values and p-values obtained, for each preprocesing method separately using plot() function of Matlab (APPENDIX H).

3.5. Zebrafish Gene Symbol Annotation

Affymetrix IDs were converted to corresponding zebrafish gene symbols using Biomart Martview service (Biomart). Affymetrix IDs were given as an input to appropriate fields and associated gene symbols were retrieved using the GUI of the web tool. 8517 genes were found corresponding to the probesets in Affymetrix Zebrafish GeneChip array.

3.6. Drawing Graphs

For the generation of boxplots and histograms, MatLab's boxplot() and hist() functions were utilized with appropriate input variables (APPENDIX D). In addition, suitable functions of AutoCAD 2008 were used for generating figures 9 and 10.

3.7. Random Network Generation

In order to compare the network topology measures with random networks, the node pairs were randomly shuffled in the upper triangle of the adjacency matrix using the code in APPENDIX E. Accordingly, given a network, the code generates a random network with the same number of nodes and edges. For comparing network results with respect to fold change and p-value distributions, actual networks were compared with random networks sampled from the original normalized dataset keeping the number of probesets the same with the intersection dataset (APPENDIX H)

3.8. Analysis of Hypoxia Dataset

To identify genes that were the least variant in both networks of different preprocessing methods, difference of a network parameter in each network was calculated and converted to its absolute value. Then, sum of the differences were sorted in order to identify the least-changed top 20% of the genes for each network topology measure. Lastly, intersection of the least-changed probeset lists was generated to show the most stable genes in all networks in terms of network topology.

CHAPTER 4: RESULTS

4.1. Effects of Preprocessing on Data Disribution

Since each preprocessing method aims to remove non-technical variation in the microarray data in different ways, they alter the distribution of the data differently. In order to show the effects of each preprocessing method, boxplots and histograms were used for visualization of each preprocessed data's numerical properties. Figure 1 demonstrates that the distribution of the raw data before any kind of preprocessing suggesting that GSE4989 arrays have comparable raw data distributions where 50% of the data resided between log2 values of 6 and 8. Raw data distribution resembled an extreme value distribution (Gumbel 1958) with a small number of probe sets with high expression values whereas close to 50% of the probes accumulated closer to the lower limit of the distribution (Figure 2). Upon preprocessing, raw data distribution has changed drastically; the shape of the distribution was dependent on the type of the preprocessing method (Figures 3, 4, and 5).

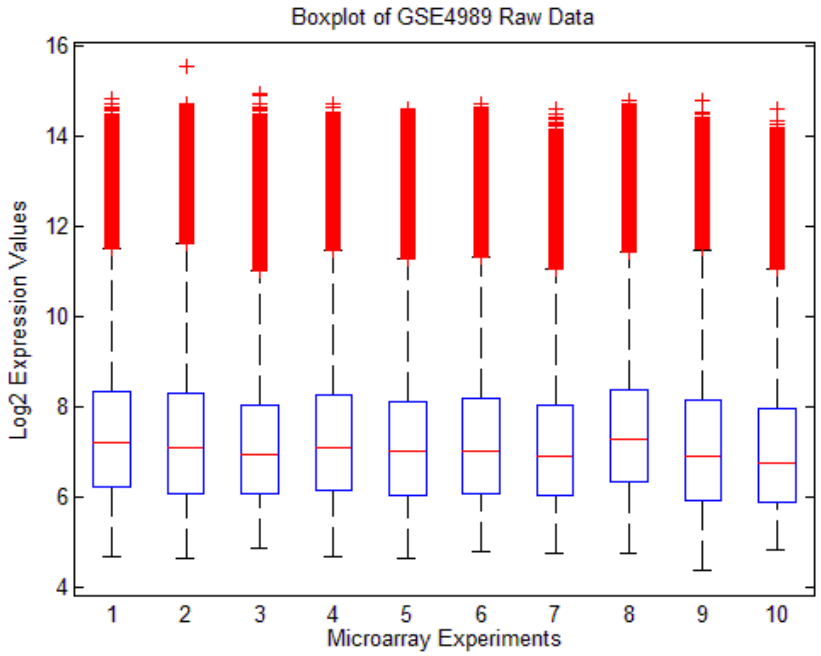


Figure 1. Data distribution of each array before preprocessing

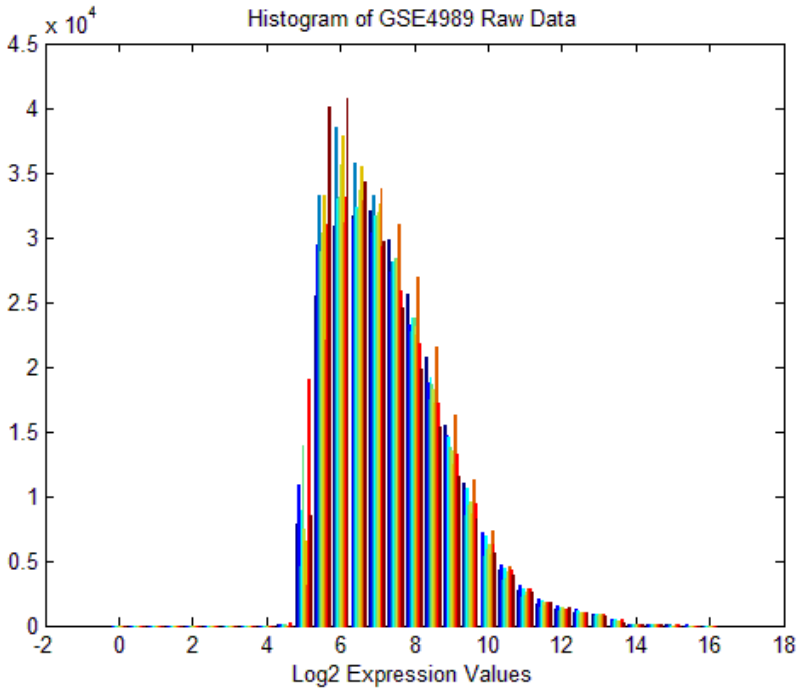


Figure 2. Raw data expression value frequencies.

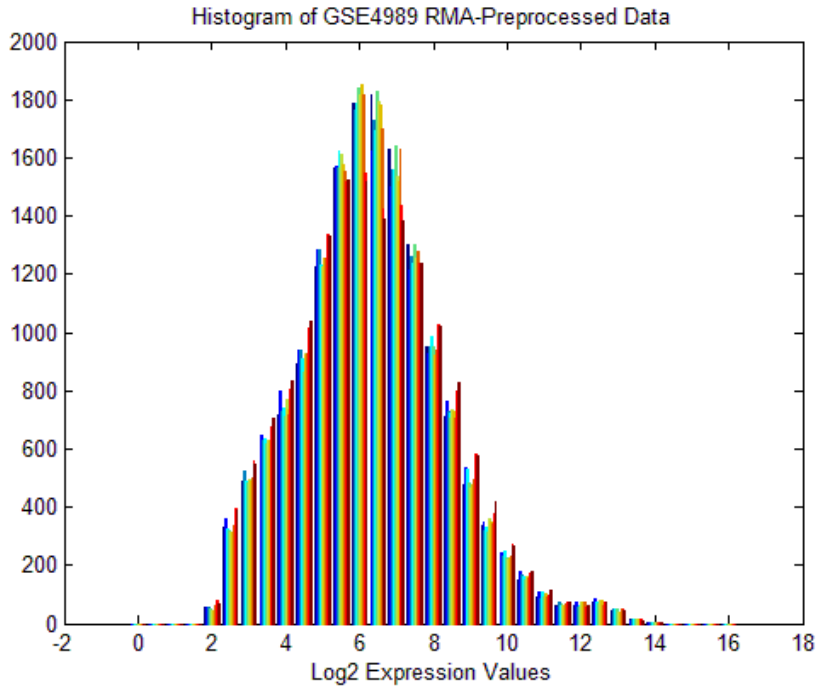


Figure 3. RMA-preprocessed data expression value frequencies.

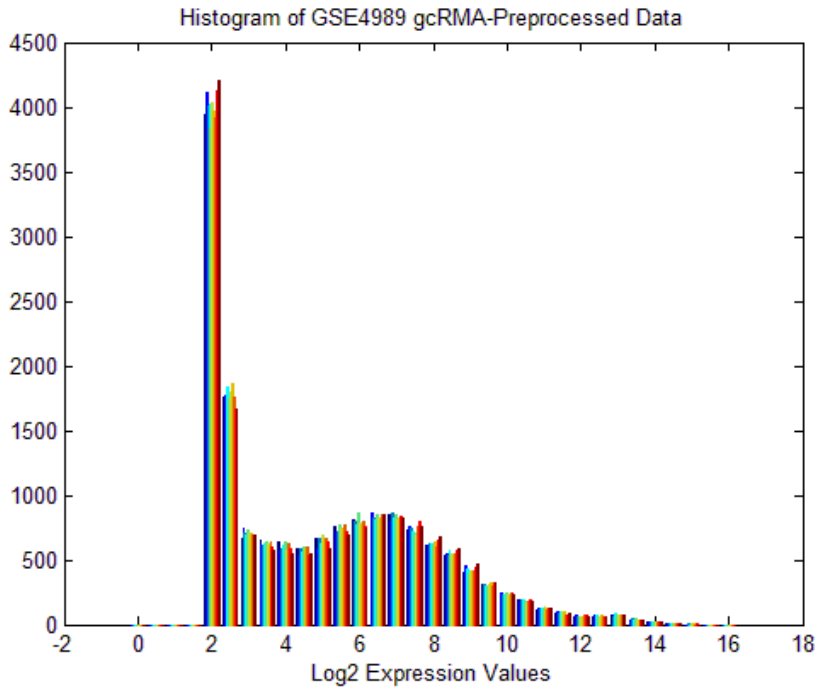


Figure 4. gcRMA-preprocessed data expression value frequencies.

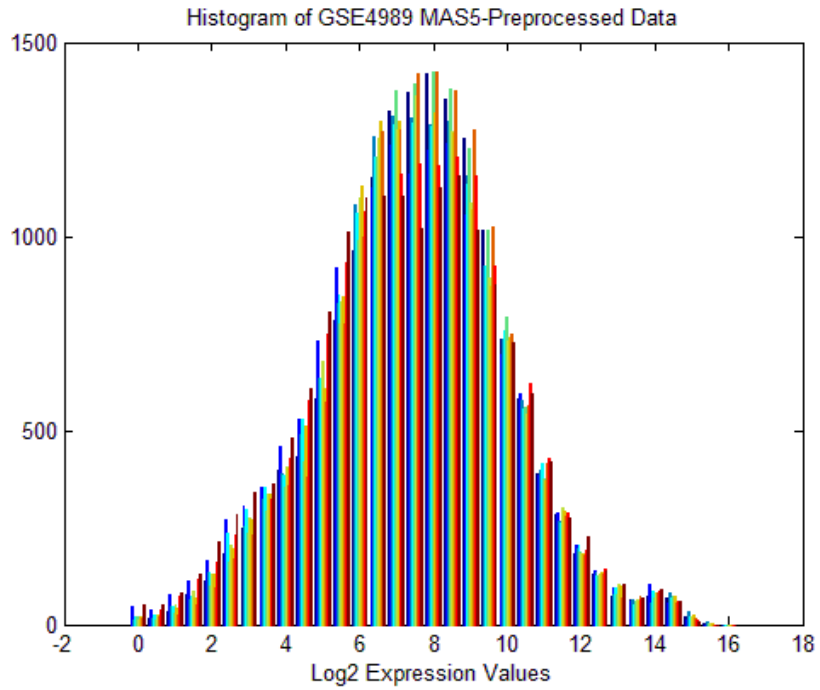


Figure 5. MAS5-preprocessed data expression value frequencies.

Based on the histogram representations, preprocessing procedures altered the data distribution. While RMA and MAS5 have produced distributions closer to normal

distribution, gcRMA led to a distribution with features of bimodality (Figure 3, 4, and 5). To be able to observe each experiment within the experiment set, boxplots were drawn. Figure 6 indicated that each method within itself was highly consistent thus arrays could be compared with each other. However, there were drastic differences with respect to the median values as well as Inter Quartile Ranges (i.e., IQRs, which is described as showing the robust 50% of the data between 25% and 75%) among different methods (Figure 6). Accordingly, MAS5 normalized experiments had higher IQRs when compared to others; the normalization method that generated the least variable experiment set was RMA. gcRMA resulted in a skewed distribution where the lower 95% confidence interval of the boxplots was truncated relative to those obtained using other methods (Figure 6).

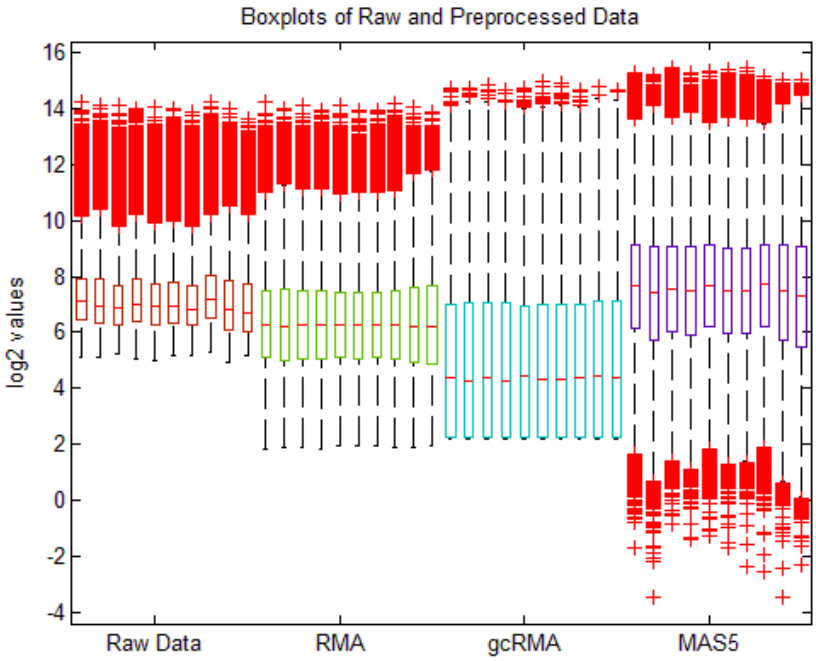


Figure 6. Comparison of distributions of raw and preprocessed data.

Different normalizations resulted in different medians (Figure 7). Also seen from the previous boxplot representation of the raw and preprocessed datasets (Figure 6), the mean of the medians from all 10 microarray experiments were less in gcRMA when compared with those from RMA and MAS5 (Figure 7). RMA data followed gcRMA data and had a 1.5 fold greater median value than gcRMA. In RMA, median values of each array seemed to be stable around the expression value of 6.3, in log2. Raw data

exhibited a higher median value compared to RMA, and lower median value compared to MAS5 (Figure 7). Since the raw data was not processed, the figure reflected the actual median expression values of each probe. In particular, the 8^{th} array seemed to have an upward shift in expression values which could be observed from the value distribution shown in Figure 6. Since RMA and gcRMA utilizes quantile normalization method normalizing each array via exchange of actual expression values between arrays, the upwards shift in the 8^{th} array was not observed in data normalized with RMA or gcRMA. Lastly, MAS5 had the highest median value for each array and also showed the widest data distribution as seen from the Figures 6 and 7. Since MAS5 used a scaling approach for preprocessing of the data, the normalization was likely to be affected by the deviation and shifts in the raw data. The median of the 8^{th} array of the MAS5 data was shifted upwards as in the raw data.

The differences in the distribution of each preprocessed data also were reflected in the standard deviation of the arrays. As seen in Figure 8, raw data exhibited the least amount of standart deviation among arrays whereas gcRMA had the highest standard deviation values. Raw data was followed by RMA and then MAS5 datasets. One clear observation from Figure 8 was that RMA and gcRMA tended to decrease the variability in distribution across arrays and standardized the distributions of each array in order to make them more comparable.

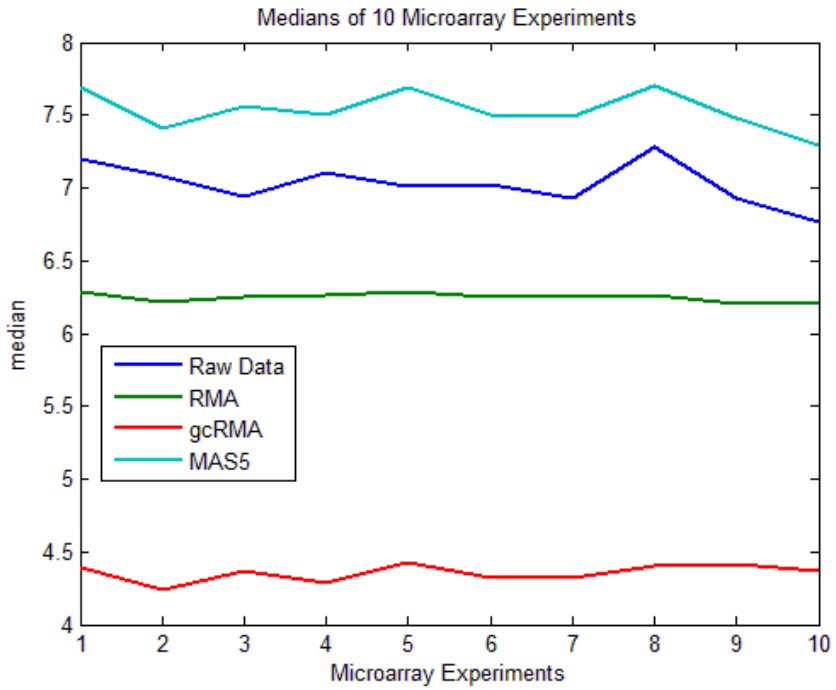


Figure 7. Plots of medians of each raw and preprocessed datasets.

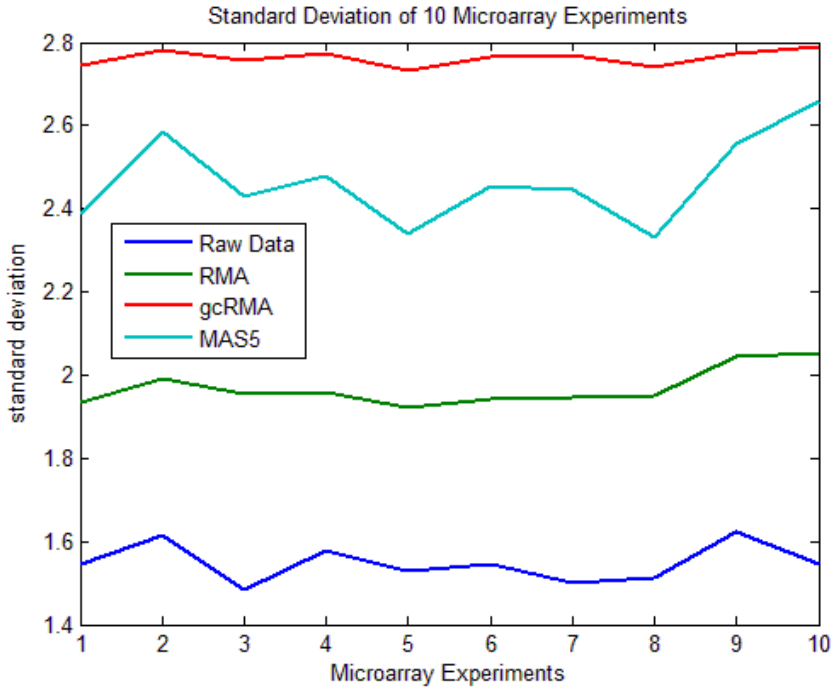


Figure 8. Plots of standard deviations of each raw and preprocessed datasets.

4.2. Effects of Preprocessing on Differently Expressed Gene Lists

Based on the findings from the histograms and boxplots of data distributions of both the raw and preprocessed data, it was clear that each preprocessing method altered the distribution and original data values in different ways. Therefore, we hypothesized that this situation also might affect the results of statistical tests based on these normalized datasets.

For each dataset, the number of differentially expressed genes was obtained using two sample equal variance t-tests with an alpha value of 0.05; and the results were represented as the number of probesets differentially expressed as well as the corresponding percentage they fell into (Tables 2 and 3). Accordingly, all three methods identified close to 20% of all probesets in the GeneChip as significant. Since gcRMA and RMA were similar in their background correction, normalization and summarization steps, the number of genes common to both methods was greater than that observed between MAS5 and RMA or MAS5 and gcRMA (Table 3). 1932 probe sets were common in these three significant gene lists, each obtained from RMA, gcRMA, and MAS5 preprocessed data, respectively (Table 3). Percentage of the common probesets among these significant gene lists was greater than 50% in all three methods (Table 4). Moreover, the union of these three gene lists consisted of 5048 unique probe sets. Union of significant gene lists of each dataset were compared and results also indicated that RMA and gcRMA gene lists were the most similar ones compared to MAS5 (Table 5). Lastly, Table 6 demonstrated the ratio of significant gene list of each dataset to the union gene list. Accordingly, MAS5 had the biggest contribution to the MAS5 list as expected.

Table 2. Number of differentially expressed genes that are detected by t-test, after each preprocessing method.

	# of probe sets	% of probe sets
RMA	3397	21.75
gcRMA	3224	20.64
MAS5	3508	22.46

Table 3. Numbers of common probe sets among pairs of differentially expressed gene lists.

Intersection	RMA	gcRMA	MAS5
RMA	3397	*2612*	*2221*
gcRMA	*2612*	3224	*2180*
MAS5	*2221*	*2180*	3508

Table 4. Percentage of common probe sets that are shared in all three differentially expressed gene lists.

Intersection	% of common probe sets
RMA	56.87
gcRMA	59.92
MAS5	55.07

Table 5. Numbers of union of differentially expressed gene list pairs.

Union	RMA	gcRMA	MAS5
RMA	3397	*4009*	*4684*
gcRMA	*4009*	3224	*4552*
MAS5	*4684*	*4553*	3508

Table 6. Percentage of differentially expressed gene lists to the union of all probe sets.

Union	% of gene lists to the union of all probe sets
RMA	67.29
gcRMA	63.86
MAS5	69.49

The results were also visualized for a better understanding of the number of probe sets that were commonly or uniquely identified by each one of the methods tested in the present study. Graphical representation and a Venn-diagram representation were

shown in the following figures (Figure 9, 10). Accordingly, each method identified a considerable number of probe sets uniquely, where MAS5 identified the most. Accordingly, gcRMA and RMA were more related to each other in terms of the number of differentially expressed probe sets.

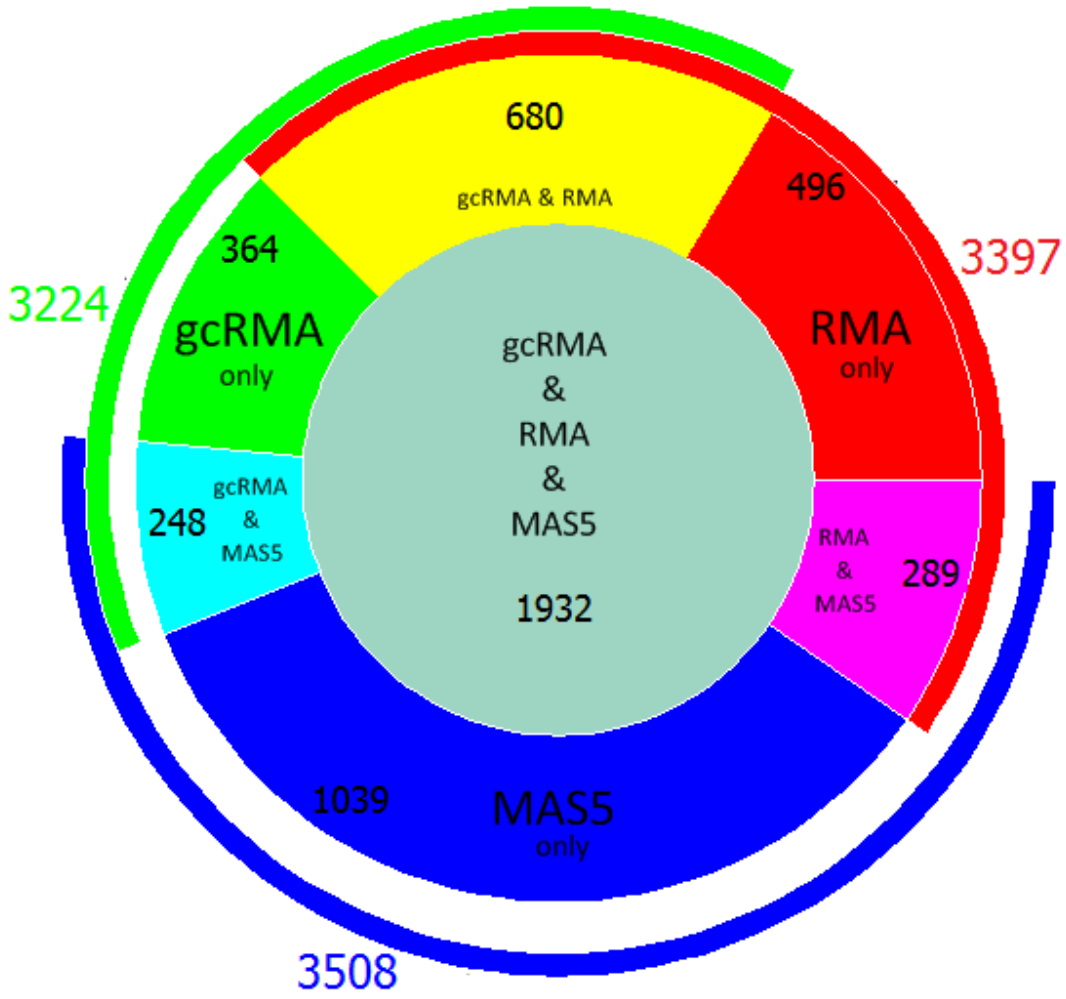


Figure 9. Distribution of intersection of probe sets generated from RMA, gcRMA, and MAS5 preprocessed data. Number of unique probe sets in each category is shown on the figure. Colored areas are proportional to the number of probe sets.

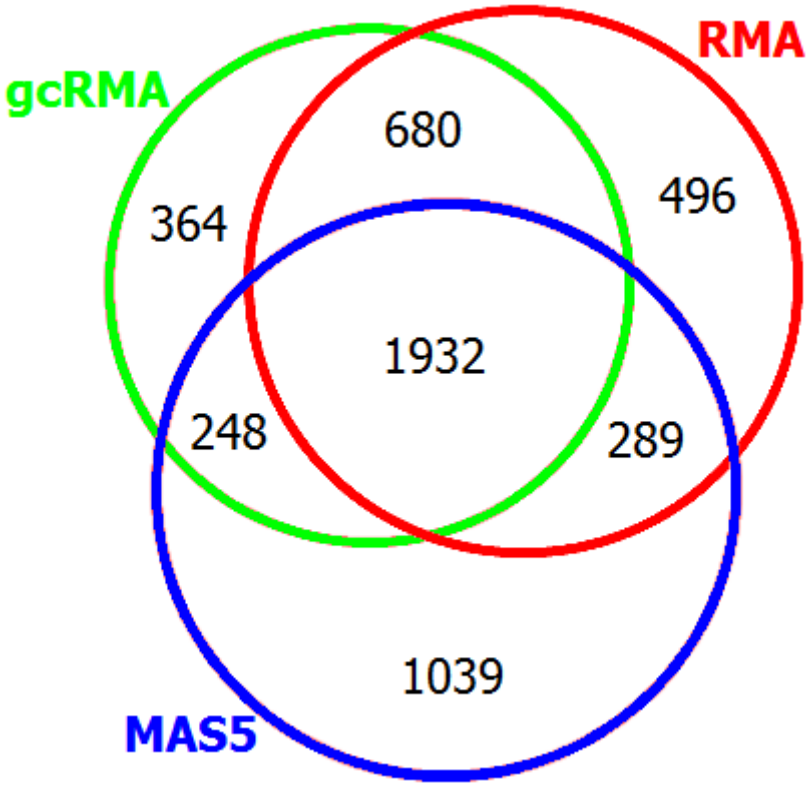


Figure 10. Venn diagram of intersection of probe sets generated from RMA, gcRMA, and MAS5 preprocessed data. Number of unique probe sets in each category is shown on the figure.

4.3. Effects of R-Value Thresholds and Preprocessing Methods on Network Generation

A landmark study performed by Lim *et al.* (2007) for investigation of the effects of normalization on gene network structure suggested of using a) arraywise correlations of real and randomized datasets; b) distribution of correlation pairs across different correlation thresholds; c) pairwise mutual information between networks; and d) functional enrichment of highly correlated pairs.

In the present study, similarly we tested whether the distribution of correlation pairs across different correlation thresholds differed according to the preprocessing method used. However, instead of analysis of arraywise correlations, we compared the average value of the positive correlation values from all pairwise probeset correlations (i.e., a mean edge correlation value). These comparisons were made between any two

preprocessing method as well as against a randomly generated network as suggested in the Lim *et al.* (2007) study. In addition, we used paired t-tests instead of mutual information indices to compare the network edge correspondence differences. Finally, frequency distribution plots of bins of correlated pairs were compared among datasets obtained from different preprocessing methods. Our analyses also made use of two different, namely the union and intersection, lists of differentially expressed genes from each method.

Probeset pairwise correlation value distributions varried from one preprocessing method to another for the union and intersection datasets (Figures 11, 12). For both datasets, correlation values were scattered to the positive and negative ends while most of the correlations were found in between 0 and -0.2. In addition, the intersection dataset had pairwise correlations accumulating at extremes of both directions suggesting a greater proportion of significant pairwise correlations. In the union data, however, except the values between 0.2 and -0.2, correlation values were relatively more uniformly distributed.

Distribution of positive correlation values ($r>0$) was visualized using boxplot representations (Figures 13, 14). Positive correlation can identify the pairs of genes having a similar expression profile among different arrays. Boxplot representations also showed a similar pattern as observed in Figures 11 and 12. Union data exhibited a more uniform distribution of correlation values whereas the median positive 'r' value was higher in the intersection data with lower interquartile of the distribution spanning a greater range of correlation values.

One additional observation from the Figures 13 and 14 was that the effect of preprocessing such that the nature of correlation could be more clearly seen in intersection data. RMA and gcRMA had similar positive correlation profiles whereas MAS5 preprocessed data had relatively lower correlation among pairs of genes. Possible reason for this difference is mentioned in Discussion section.

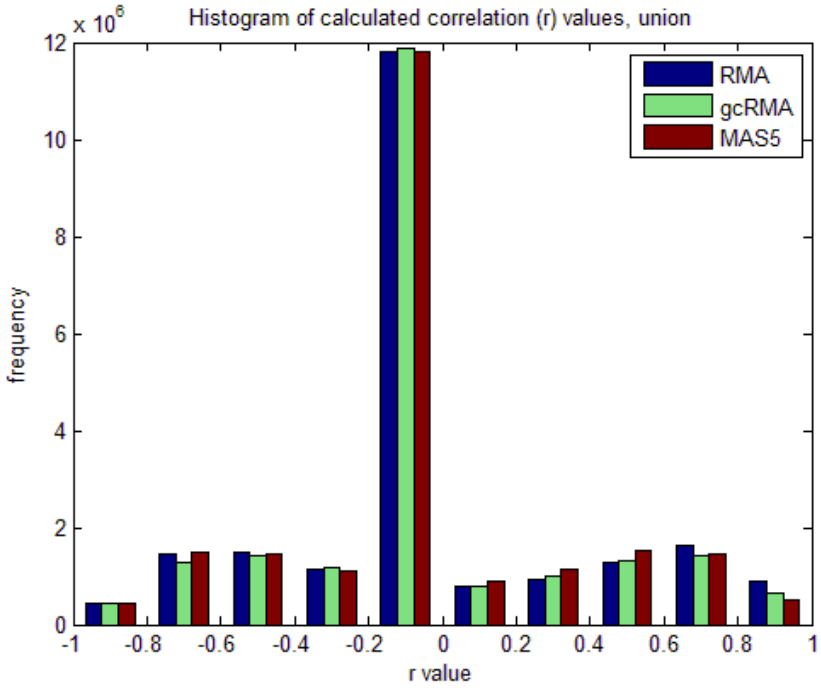


Figure 11. Histogram of correlation values from union data of each preprocessing method.

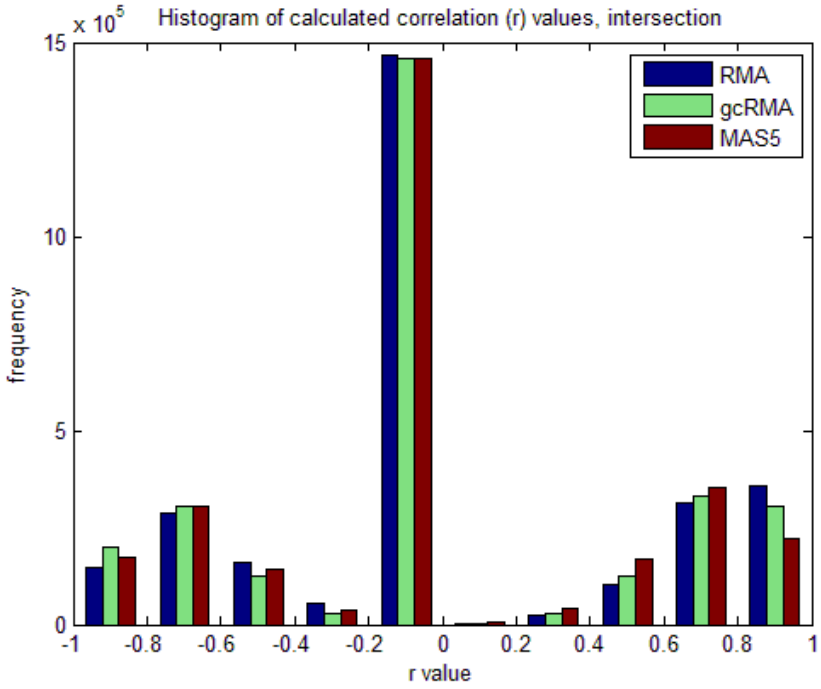


Figure 12. Histogram of correlation values from intersection data of each preprocessing method.

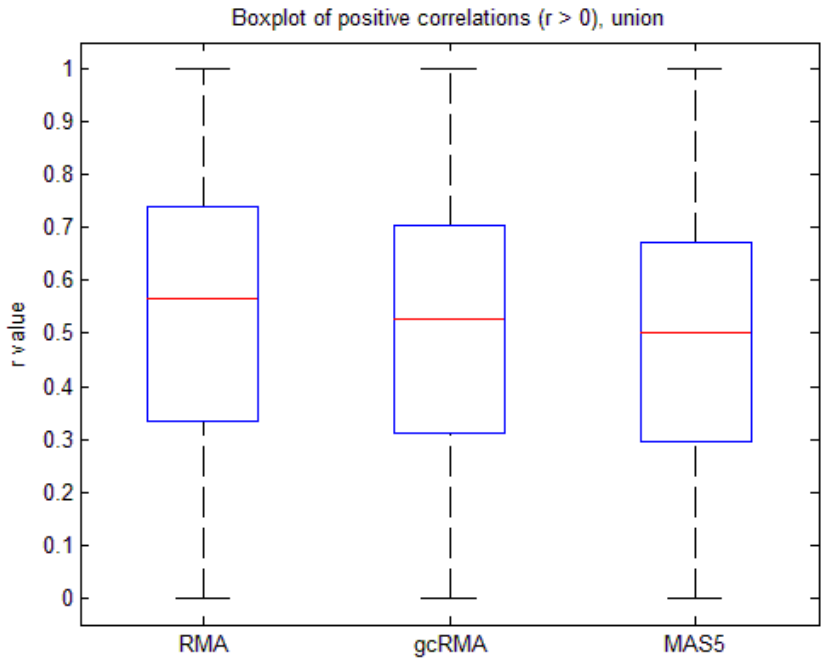


Figure 13. Distribution of positive correlation values in each preprocessed union data

.

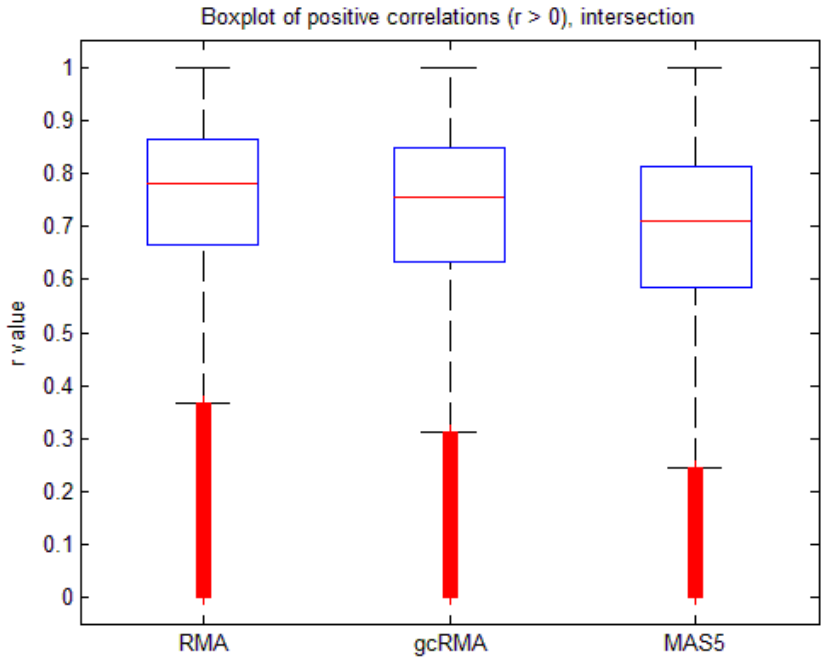


Figure 14. Distribution of positive correlation values in each preprocessed intersection data.

To demonstrate the effects of varying the r-value threshold on the number of edges in each network, distribution of sum of all gene correlations above the threshold (edges) were plotted separately for the union and intersection data (Figures 15, 16). These graphs showed that although there was a slight difference between the correlation distributions of differently preprocessed data, choosing an r-value of 0.6 and greater could minimize the nonlinearity due to methodology. For example, the distribution of all three methods ran parallel to each other when r equaled to or was greater than 0.6. Although RMA- and gcRMA-based correlation values had similar slopes, MAS5-based correlation values decreased faster as the r-value increased (Figure 15). The r-value around 0.45 was critical in this context so that MAS5-based data became the least correlated in terms of the number of gene pairs. Since intersection data had more significant genes having similar gene expression profiles, MAS5 remained the least correlated data however the slopes were parallel exhibiting linearity among preprocessing methods (Figure 16).

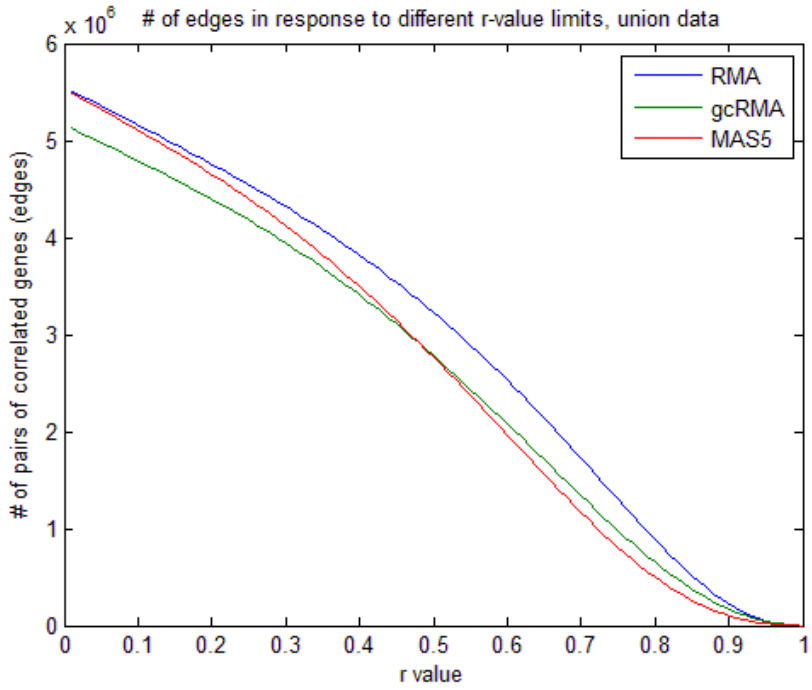


Figure 15. Sum of edges for networks generated at different r-value thresholds for union data.

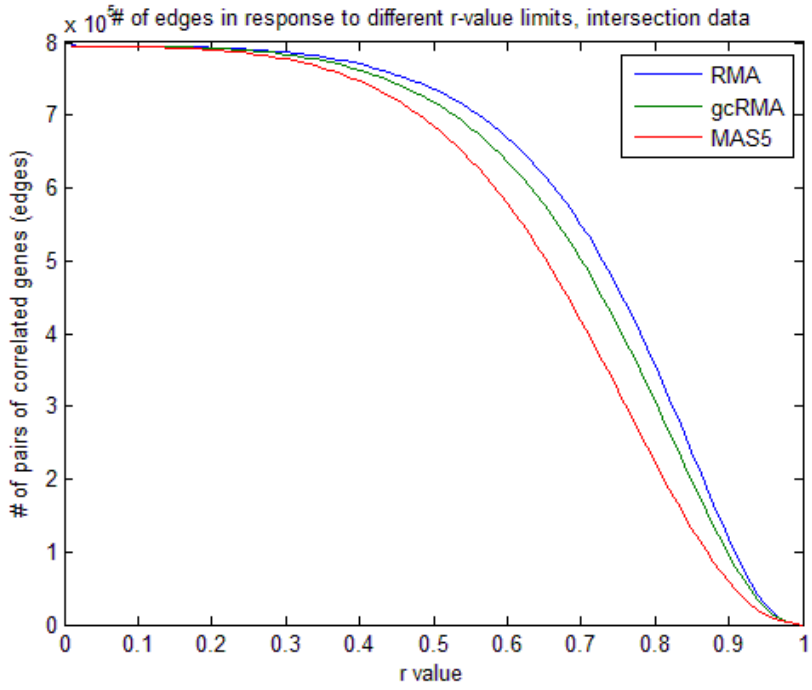


Figure 16. Sum of edges for networks generated at different r-value thresholds for intersection data.

4.4. Effects of preprocessing methods on network structure

To assess the effects of preprocessing methods on network structure, three network topology measures have been widely utilized: betweennes centrality, clustering coefficient, and the degree distribution or connectivity (Barrat *et al.* 1999; Freeman 1977; Newman 2003). Indeed, these three important network parameters also were previously used in comparison of networks generated from protein-protein interaction, radiation hybrids, functional annotation, and gene expression datasets (Ahn *et al.* 2009).

4.4.1. Betweenness centrality

Genes located among the shortest paths between any other two genes have a higher betweenness centrality value. Thus, genes with a higher betweenness centrality are thought to have a central role for cellular functions especially having roles for communication between modules (Hintze *et al.* 2008).

Distribution of betweenness centrality was calculated for correlation networks of each preprocessed data for union (Figures 17-18) and intersection (Figure 19) data. Each preprocessing method was analyzed separately, a boxplot was drawn showing the distribution of the betweenness centrality value distribution when compared with that from a random distribution. Network of original data and random data distribution significantly differed with respect to the range of distribution (Figure 17-19); for both the union and intersection data, RMA, gcRMA, and MAS5 betweenness centrality values were significantly different from their random counterparts with a p-value of zero (less than 10E-16). The random networks were designed to have the same number of genes and pairwise edges, with a randomized correlation distribution. Compared to the random networks, each preprocessed original data had a wide spectrum of betwenness centrality values expected to be seen in a real situation.

On the other hand, preprocessing methods did not differ among each other with respect to betweenness centrality data distribution based on the paired-t-tests (Table 7) in the union data. However, usage of intersection data accentuated the differences among the preprocessing methods such that there was an increase in the median and IQR of the between centrality measurements obtained via MAS5 normalization (Table 8). These might show that the different preprocessing methods affected the highly significantly expressed genes in a network setting.

Table 7. Comparisons of the betweenness centrality distributions of RMA, gcRMA, and MAS5 preprocessed correlation networks using one-sampled t-test, for union data.

Betweenness Centrality	**gcRMA**	**MAS5**
RMA	0.6167	0.0923
gcRMA		0.3536

Table 8. Comparisons of the betweenness centrality distributions of RMA, gcRMA, and MAS5 preprocessed correlation networks using one-sampled t-test, for intersection data.

Betweenness Centrality	gcRMA	MAS5
RMA	0	0
gcRMA		0

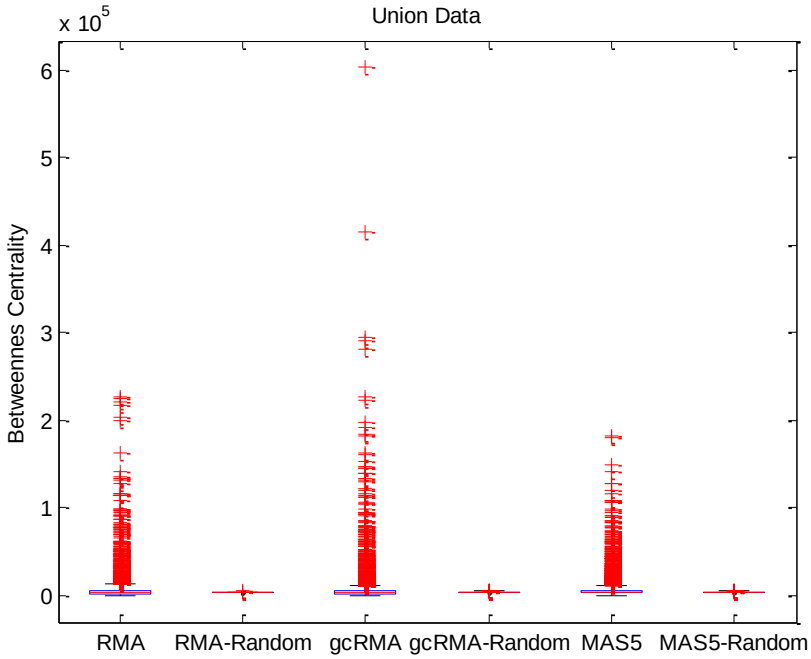


Figure 17. Boxplots representing the distribution of betweenness centrality values in each network for union data.

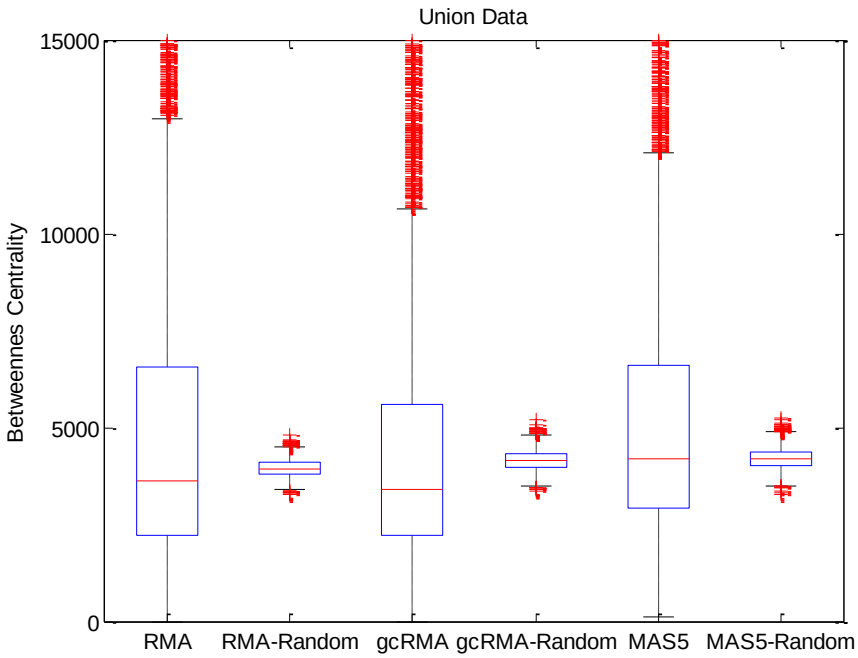


Figure 18. Detailed representation of Figure 17 for better visualization of the distributions between the first and the third quarter.

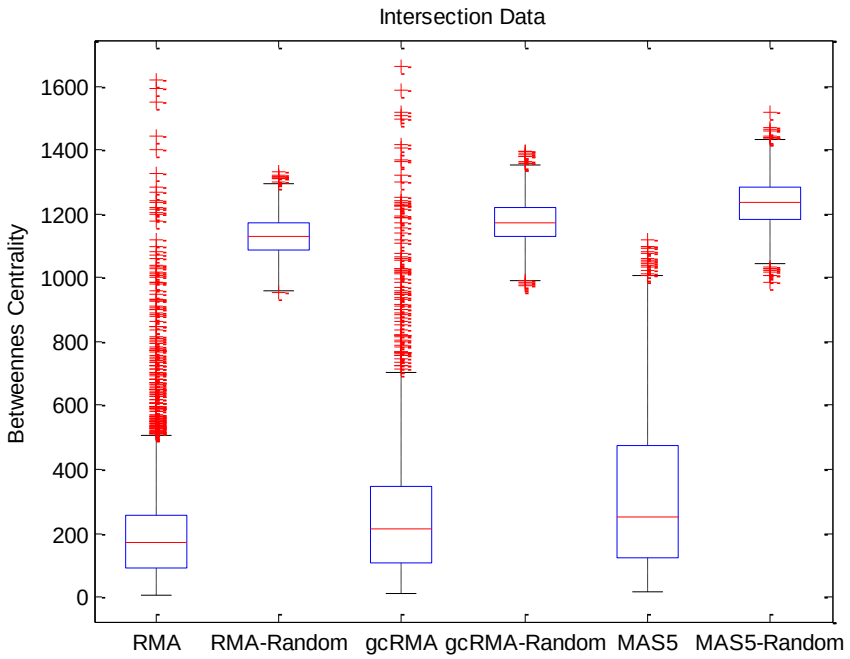


Figure 19. Boxplots representing the distribution of betweenness centrality values in each network for intersection data.

4.4.2. Clustering coefficient

Clustering coefficient is a network topology parameter for measuring the clustering tendency of nodes/genes. Lower clustering coefficients are the indicators of random networks (de Haan *et al.* 2009). A gene with a higher clustering coefficient is thought to have an actively interacting profile with other genes (Horvath *et al.* 2008).

To assess the clustering tendency of each network, the clustering coefficient distributions were calculated; boxplotted; and compared with their random networks in pairs (Figures 20-21). Results of comparisons with random networks gave similar results with those of betweenness centrality; clustering coefficients of networks of actual data were significantly different from random counterparts with p-values of zero (less than 10E-16). Strikingly, random networks exhibited very low clustering coefficient values as expected. When the clustering coefficients of RMA, gcRMA, and MAS5 networks were compared in a pairwise fashion, it was observed that clustering coefficient distributions of each preprocessing network was significantly different from each other for both the union and intersection data (Tables 9-10). This

result indicated that the clustering tendency of differentially expressed genes' were highly affected by the nature of the preprocessing method. Although RMA and gcRMA had very similar protocols for preprocessing, the correction for the GC bias in probesets seemed to have an affect the network structure significantly.

Table 9. Comparisons of clustering coefficient distributions of RMA, gcRMA, and MAS5 preprocessed correlation networks using one-sampled t-test, for union data.

Clustering coefficient	gcRMA	MAS5
RMA	0	0
gcRMA		0

Table 10. Comparisons of clustering coefficient distributions of RMA, gcRMA, and MAS5 preprocessed correlation networks using one-sampled t-test, for intersection data.

Clustering coefficient	gcRMA	MAS5
RMA	0	0
gcRMA		0

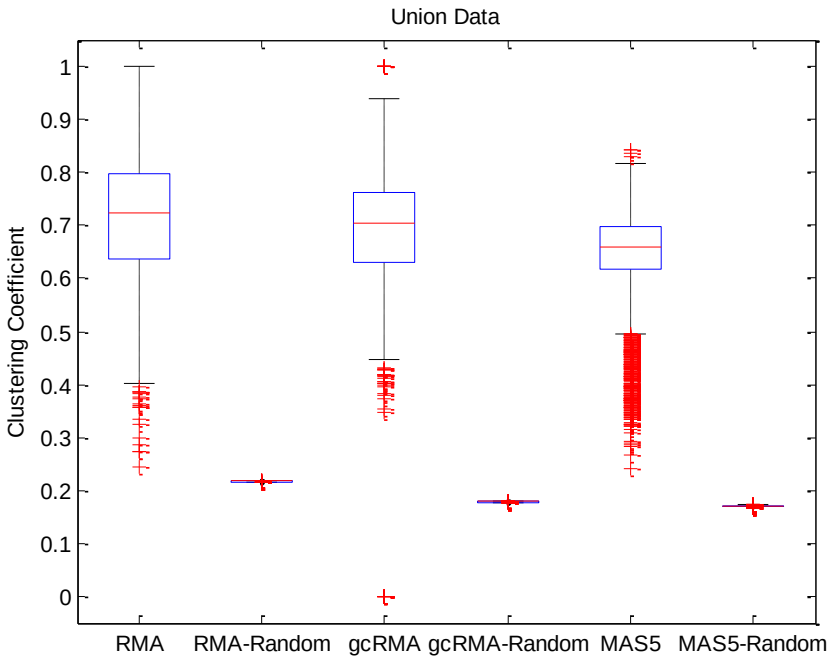


Figure 20. Distributions of clustering coefficients among different networks, for union data.

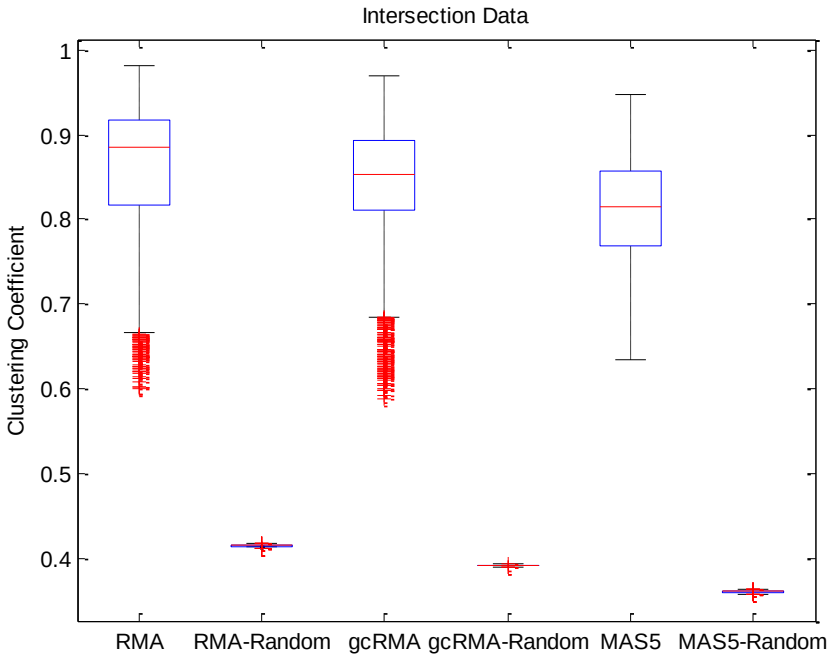


Figure 21. Distributions of clustering coefficients among different networks, for intersection data.

4.4.3. Degree distribution

Degree distribution or connectivity of a network is another measure for understanding the dynamics of a network. Degree of a node/gene shows the number of its correlated pairs of genes. Connectivity has been shown to be correlated with essentiality of gene function (Caretta-Cartozo *et al.* 2007).

Following figures helped visualize the effects of preprocessing methods on connectivity (Figures 22-23). When compared with the random networks, actual networks were not significantly different (p-values greater than 0.99) in terms of median values for both the union and intersection data. However, when IQR of the actual and random networks were considered, random networks have a much more uniform distribution of node-degree (Figures 22-23). Pairwise comparisons of actual networks were significantly different with very low p-values indicating that the network structure was highly affected in terms of the number of correlations (Tables 10-11). gcRMA and MAS5 resulted in a decrease in nodes with greater number of edges when compared with RMA especially for the intersection dataset (Table 12; Figure 23). Interestingly, random networks also exhibited similar declines in node-degree in the same direction, e.g., RMA>gcRMA>MAS5. Median connectivity was greater in the random network in comparison with the real in RMA and gcRMA, Strikingly, in MAS5 median connectivity of the random network was less than that of the real network.

Table 11. Comparisons of degree distributions of RMA, gcRMA, and MAS5 preprocessed correlation networks using one-sampled t-test, for union data.

Degree distribution	gcRMA	MAS5
RMA	0	0
gcRMA		06.0209e-006

Table 12. Comparisons of degree distributions of RMA, gcRMA, and MAS5 preprocessed correlation networks using one-sampled t-test, for intersection data.

Degree distribution	gcRMA	MAS5
RMA	0	0
gcRMA		0

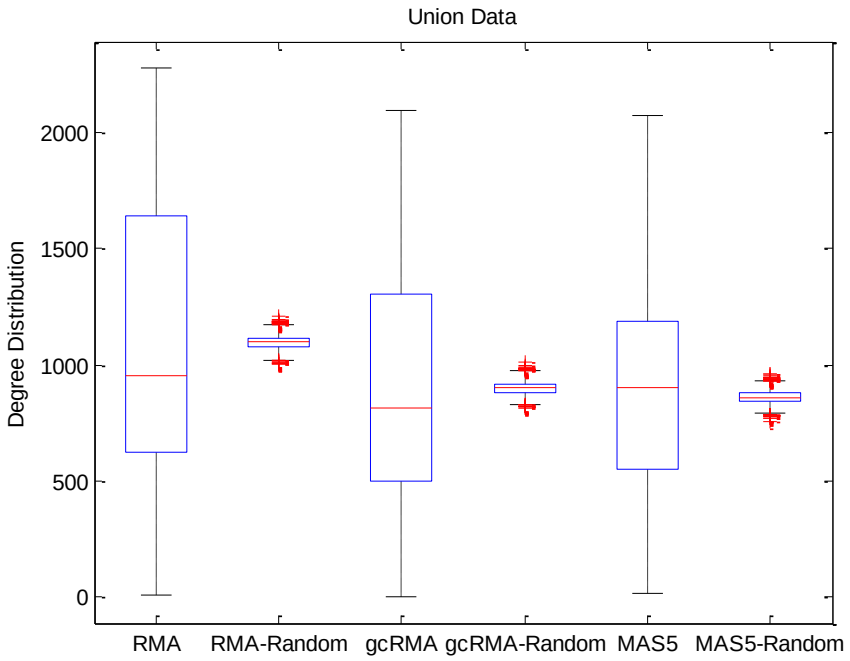


Figure 22. Degree distribution in different networks, calculated for union data.

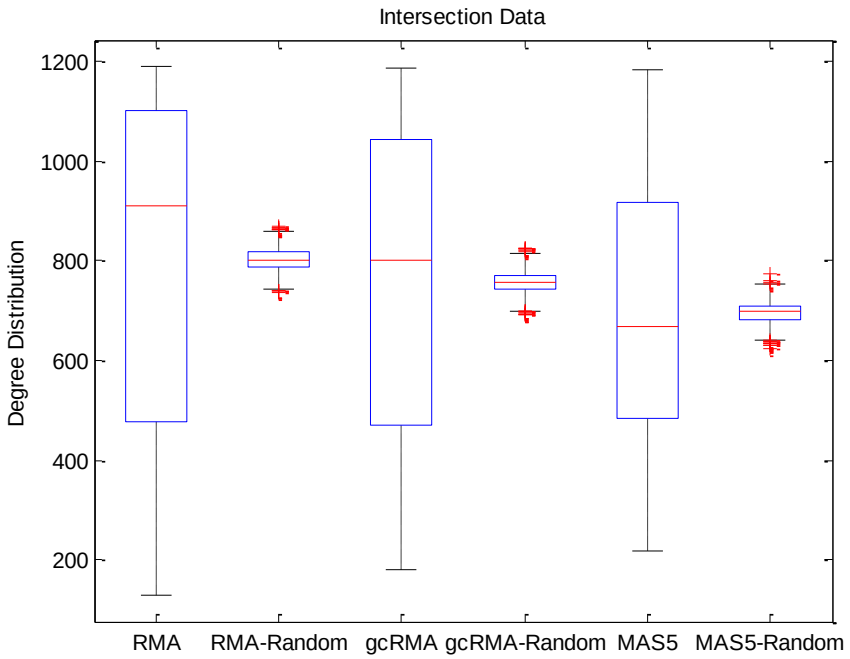


Figure 23. Degree distribution in different networks, calculated for intersection data.

4.5. Up- and Down-regulated Probeset Network Structure Comparisons

Comparisons between actual and random networks indicated that intersected data might reflect the differences among preprocessing methods better since all the nodes in this network were significantly regulated in hypoxia in zebrafish. The networks were also characterized by only the positive correlations thus representing the upregulated and downregulated probesets in hypoxia, which might result in separation of network into two accordingly. To test whether upregulated and downregulated probesets exhibited differences in their betweenness centrality, degree distribution, and clustering coefficients, we used fold change and p-values as indicators of network classification. In the literature, analysis of upregulated and downregulated gene networks represent a promising area of research (Hernández *et al.* 2007; Swindell 2008; Wachi *et al.* 2005) suggesting distinct differences inbetween.

In the present thesis as well, the upregulated and downregulated genes showed different network characteristics for the selected network measures. In the Figure 24, the difference in terms of clustering coefficient was apparent when it was plotted against the values of the fold change. It was observed that clustering coefficients were

higher on average in the downregulated genes of RMA and gcRMA networks. However, MAS5 data did not exhibit such a strong difference. In addition, in terms of the cluster of the clustering coefficient measure, it was seen in both gcRMA and RMA datasets that upregulated genes were more scattered in their range than the downregulated ones.

Furthermore, clustering coefficients were also plotted against the p-value of each probeset in the intersection data (Figure 25). The results indicated that there was a non-linear decrease in the clustering coefficients as the p-values got more significant for both the upregulated and downregulated probesets. Although the trajectories resembled each other, the downregulated probesets (represented by blue dots) had relatively higher clustering coefficients in comparison with upregulated probesets (represented by red dots) for RMA and gcRMA. However, MAS5 normalized data exhibited such a difference only for the genes with relatively low p-values thus more significantly differentially expressed (Figure 25). As the p-values became higher, the difference between upregulated and downregulated probeset distributions in terms of clustering coefficient decreased.

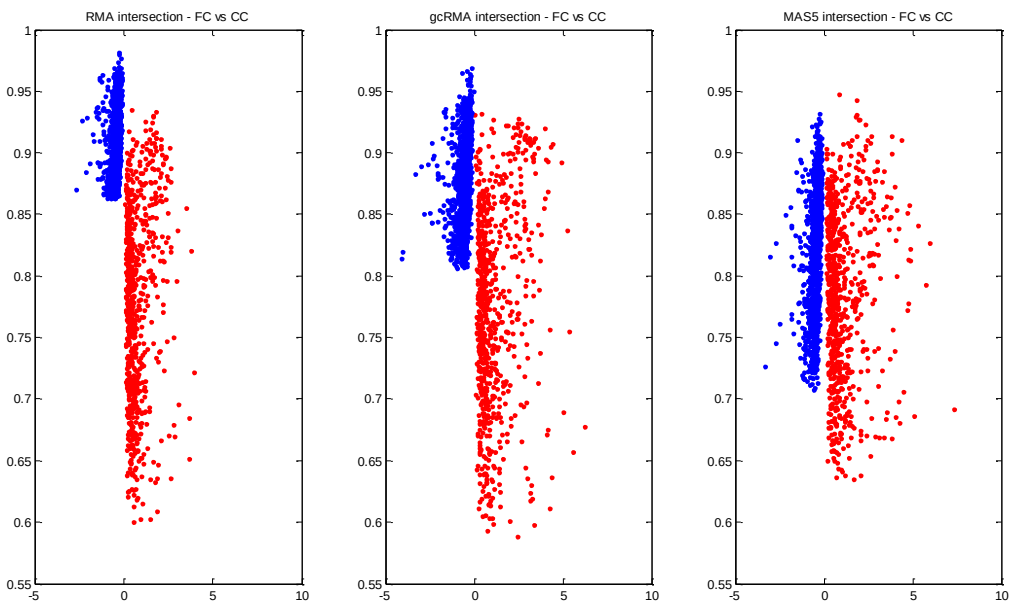


Figure 24. Fold change versus clustering coefficient for the networks of intersection data. Red dots represent upregulated genes.

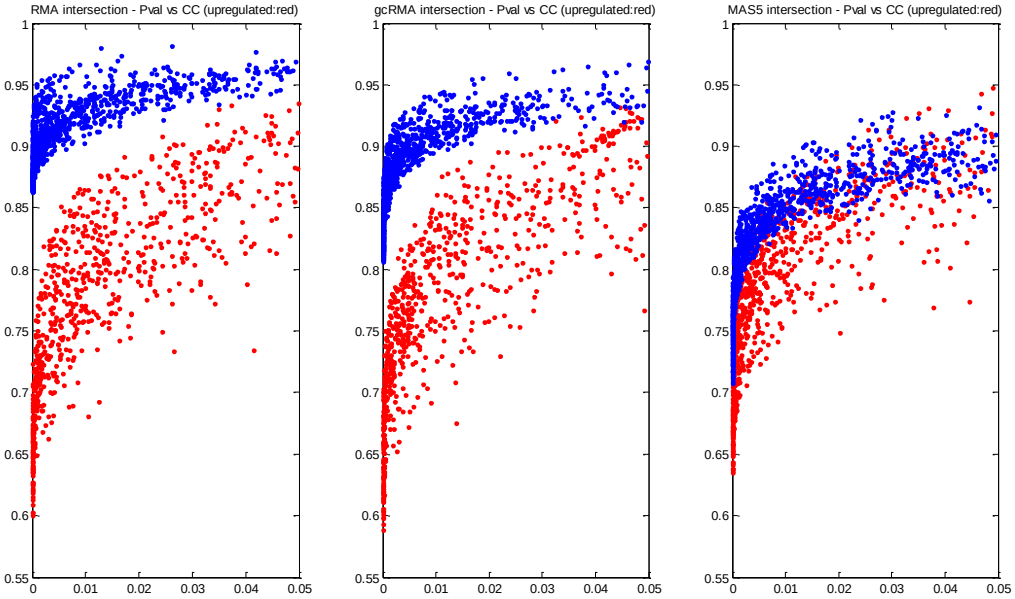


Figure 25. P-value versus clustering coefficient for the networks of intersection data. Red dots represent upregulated genes.

The difference between upregulated and downregulated genes was also observed for betweenness centrality (Figure 26). RMA and gcRMA networks had relatively low betweenness centrality values whereas MAS5 has a similar maximum and minimum values for both upregulated and downregulated genes. However, downregulated genes were clustered in terms of betweenness centrality in gcRMA and RMA preprocessed dataset when compared with upregulated genes, which were scattered across a greater range of network parameter measurement.

Similar differences among preprocessing methods were observed when betweenness centrality was plotted against the range of p-values in the intersection dataset (Figure 27). Downregulated genes had lower values in RMA and gcRMA data. Also, the difference between the distrubutions of the upregulated and downregulated genes was highly accentuated in the lower p-value intervals. This difference could not be observed in MAS5 network even for the low p-values.

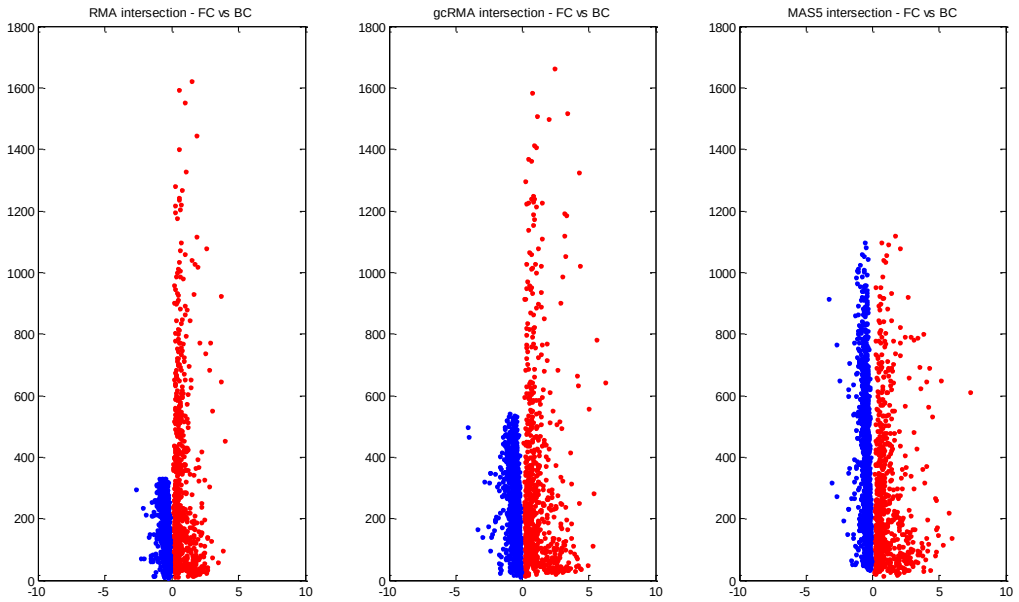


Figure 26. Fold change versus betweenness centrality for the networks of intersection data. Red dots represent upregulated genes.

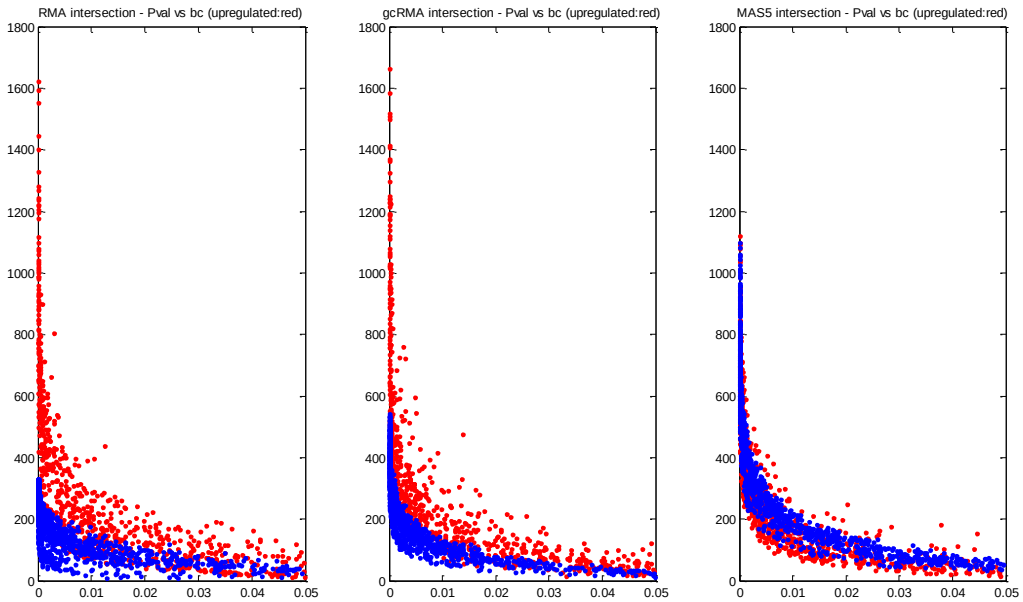


Figure 27. P-value versus betweenness centrality for the networks of intersection data. Red dots represent upregulated genes.

Lastly, degree distribution versus fold change graphs were plotted in Figure 28. Strikingly, upregulated genes were less correlated among themselves in comparison with the downregulated genes. Major difference between RMA versus gcRMA and MAS5 was that the network of MAS5 and to a lesser degree that of gcRMA contained downregulated genes spanning a greater range of node-degrees for the down regulated genes (Figure 28). This might be the reason behind the relatively higher clustering coefficients in downregulated genes in RMA and gcRMA. Upregulated genes had similar boundaries among different networks whereas gcRMA and MAS5 had greater scatter across the fold change values.

Interestingly, in the case of degree distribution, all three preprocessing methods showed similar patterns with respect to up- and down-regulated genes across p-values; genes with lower p-values tended to have more degrees suggesting that they had more correlated pairs compared to the rest. Downregulated genes had higher degrees which might explain the clustering tendency among those genes in terms of clustering coefficient. Although RMA network had a more scattered distribution of values compared to gcRMA and MAS5. Moreover, this network had a lower slope among downregulated genes as the p-value increased.

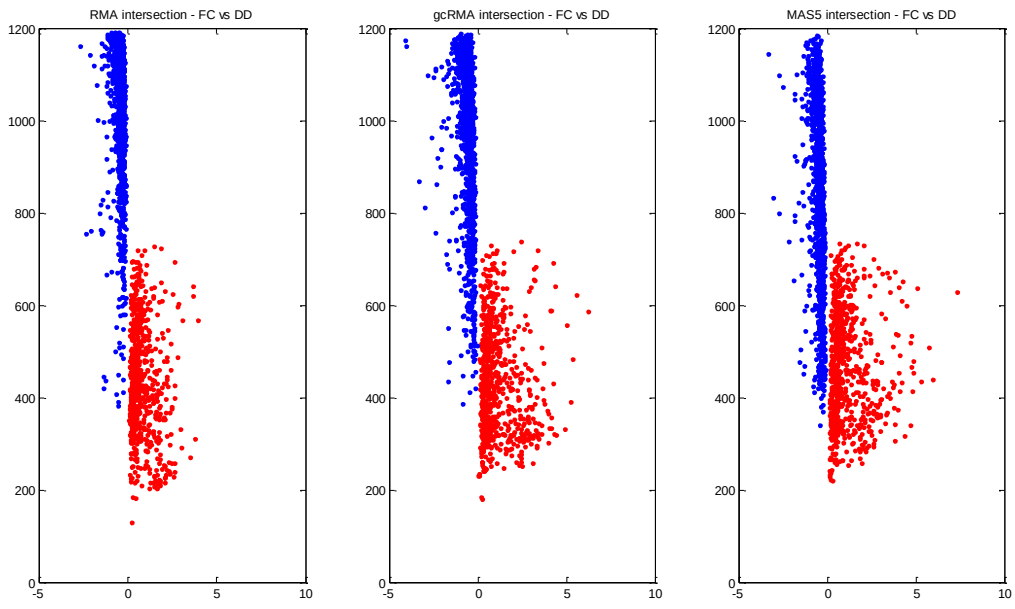


Figure 28. Fold change versus degree distribution for the networks of intersection data. Red dots represent upregulated genes.

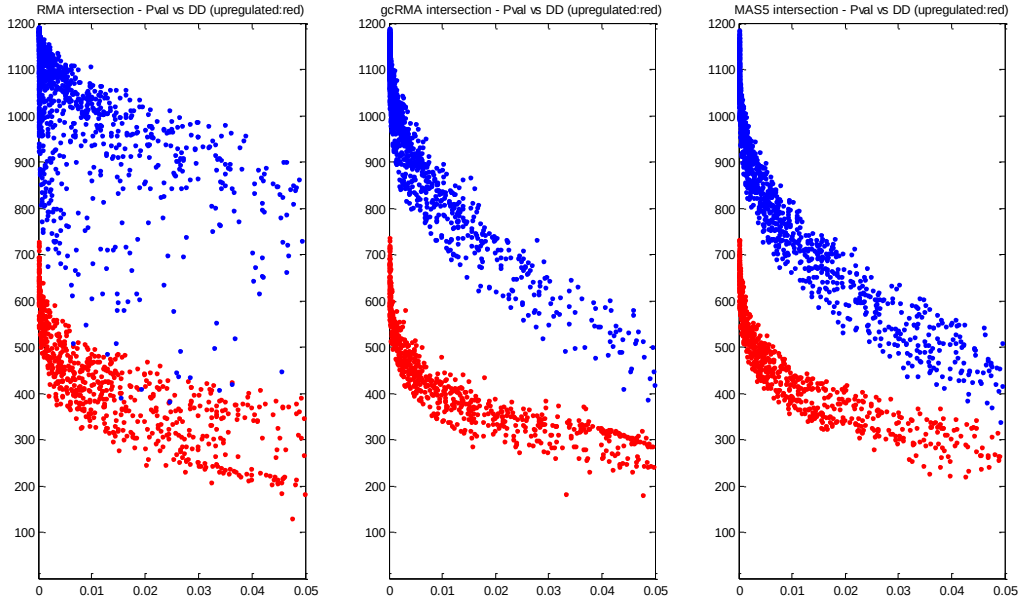


Figure 29. P-value versus degree distribution for the networks of intersection data. Red dots represent upregulated genes.

To get a better comparison of each network topology measure among methods, scatter plots among pairs of different preprocessing methods were plotted as used by Ahn *et al.* (2009). Common to the following figures (Figures 30-32) was that, RMA and gcRMA had very similar network topology measures since the scatter was located mostly diagonally. However, comparisons with MAS5 indicated a higher scatter, a sign of lower similarity. In terms of gene regulation, upregulated genes were highly conserved among different networks whereas downregulated genes were more sensitive to the preprocessing method. This was more easily observed in the scatter plot of degree distribution in Figure 32. Also, Spearman correlation coefficients for the comparison of each network topology measure for each preprocessing dataset were shown in Table 13. It is observed that RMA is more similar to gcRMA, especially for clustering coefficient.

Each scatter plot was characterized with a correlation coefficient and summary of the rho values calculated from the spearman correlation is given in the table below:

Table 13. Spearman correlation rho values of pairwise comparison of each network topology measure matrix

	Clustering coefficient		Betweenness centrality		Degree distributon	
	gcRMA	**MAS5**	**gcRMA**	**MAS5**	**gcRMA**	**MAS5**
RMA	0.88	0.56	0.79	0.55	0.94	0.82
gcRMA		0.61		0.59		0.83

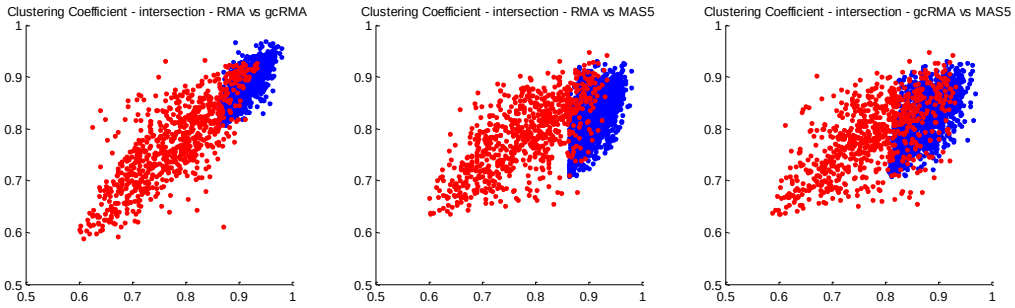


Figure 30. Scatter plots of clustering coefficients for each network pair. Red dots represent upregulated genes.

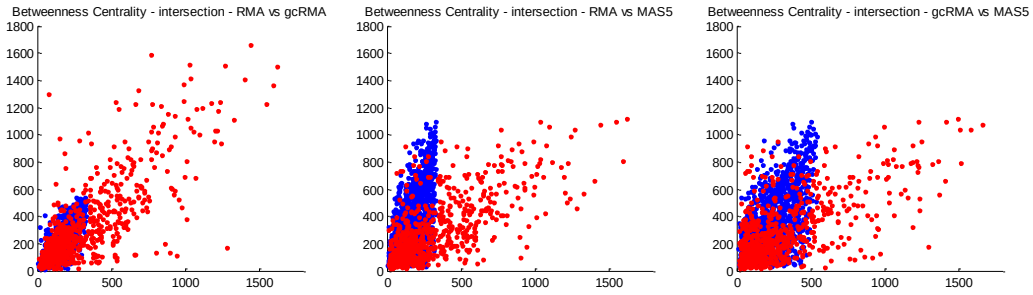


Figure 31. Scatter plots of betwenness centrality values for each network pair. Red dots represent upregulated genes.

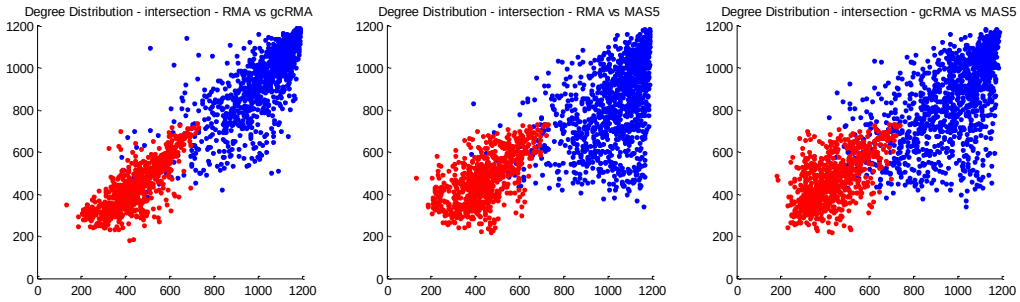


Figure 32. Scatter plots of degree distribution values for each network pair. Red dots represent upregulated genes.

4.6. Comparison of Network Topology Measures between the Real and Randomly Generated Networks

The zebrafish GeneChip contains 15617 probesets, 1932 of which were contained in the intersection dataset. These probesets exhibited significant differences under hypoxia (down- or up- regulated). Therefore, the analyses explained in Section 4.5 reflected the network structure and its variation with respect to fold change and p-value distritutions of a highly significant gene list. A random sampling of the same data size as the intersection dataset was performed to visualize how the network parameters from a network containing probesets with p-values ranging between 0 and 1 behaved with respect to fold change and p-values. The corresponding networks for each RMA, gcRMA, and MAS5 preprocessed data were generated for the random selection; the network topology measures for each network were calculated and corresponding graphs were plotted in the following figures.

First, the relationship between fold change and clustering coefficient was investigated in the random data. As seen in Figure 33, the previously observed pattern of the difference among upregulated and downregulated genes in RMA and gcRMA was not apparent. In figure 34, this situation was more clearly seen such that clustering coefficient was scattered mostly when p-value was lower than 0.05. This might mean that non-significant genes did not show a pattern for either of the preprocessing datasets.

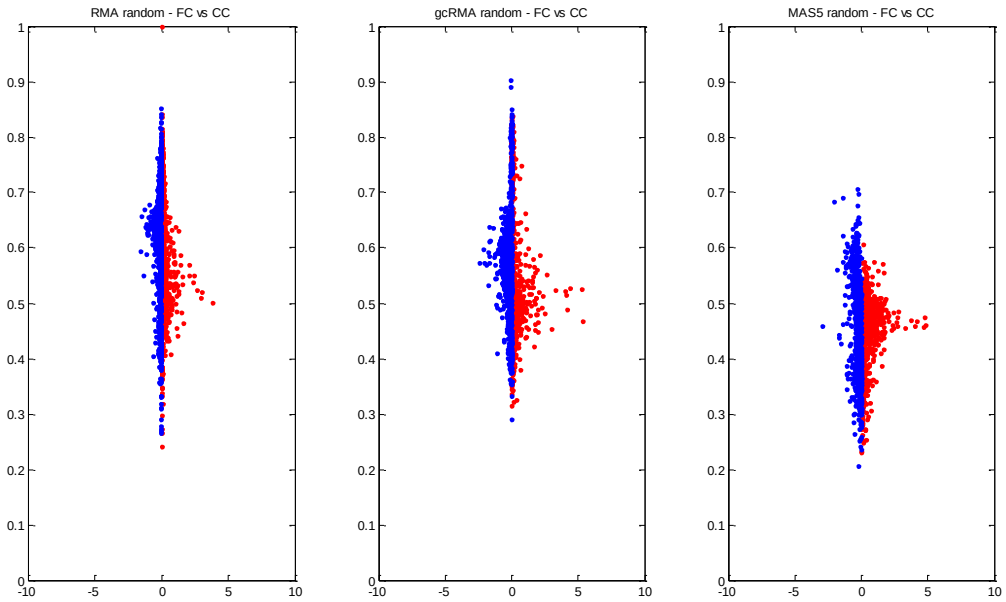


Figure 33. Fold change versus clustering coefficient for the networks of random data. Red dots represent upregulated genes.

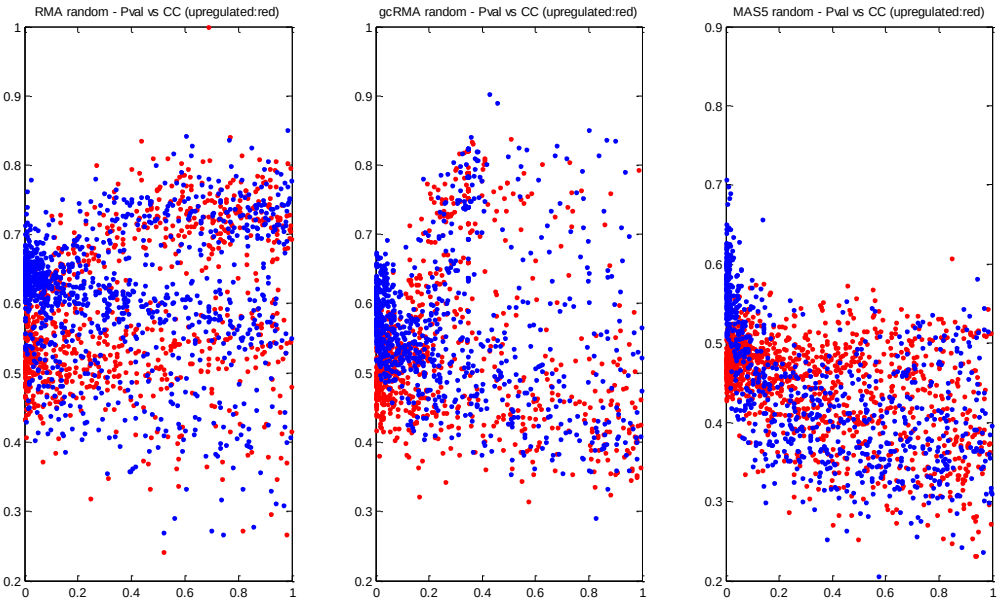


Figure 34. P-value versus clustering coefficient for the networks of random data. Red dots represent upregulated genes.

Betweenness centrality when plotted against fold change was symmetrically distributed around 0 in random data (Figure 35) unlike the intersection data (Figure 26). In contrast to the intersection networks, networks of the random data seemed to be scattered randomly when betweenness centrality was plotted against the p-value (Figure 36). In addition, there was no cluster of genes in terms of up or downregulation.

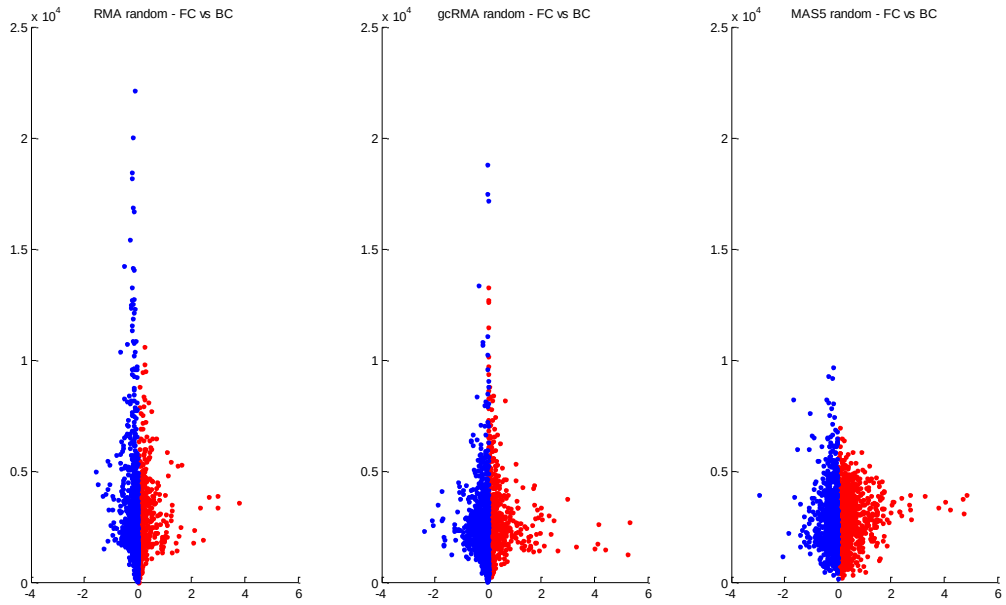


Figure 35. Fold change versus betweenness centrality for the networks of random data. Red dots represent upregulated genes.

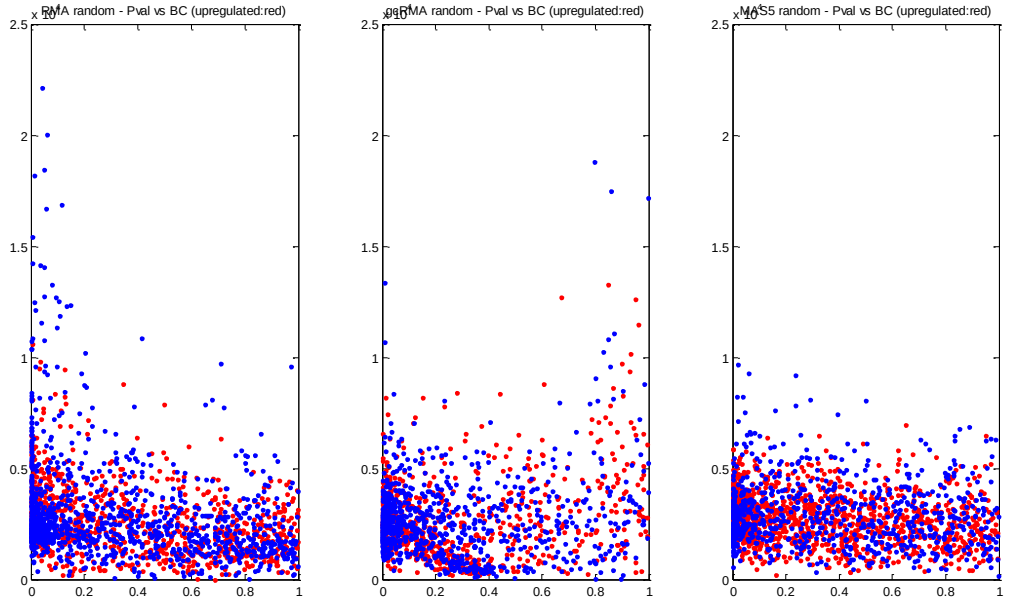


Figure 36. P-value versus betweenness centrality for the networks of random data. Red dots represent upregulated genes.

In addition to the previous network topology measures, degree distribution was investigated in the networks of random data. Although node-degrees of downregulated genes were higher in intersection data (Figure 28), networks of the random data showed a different pattern from one preprocessing method to another (Figure 37). From the point of view of p-value versus degree distribution, there again was no consensus among the networks of different preprocessing methods (Figure 38). Although RMA network seemed to have a random distribution of node-degree versus p-value, the upregulated and downregulated genes clustered in gcRMA and MAS5 graphs. In addition, degree distribution of non-significant genes did not show a consistent profile among different networks of preprocessing methods. This might show the sensitivity of the node-degree to different normalization methods where there is no significant regulation in a cellular condition.

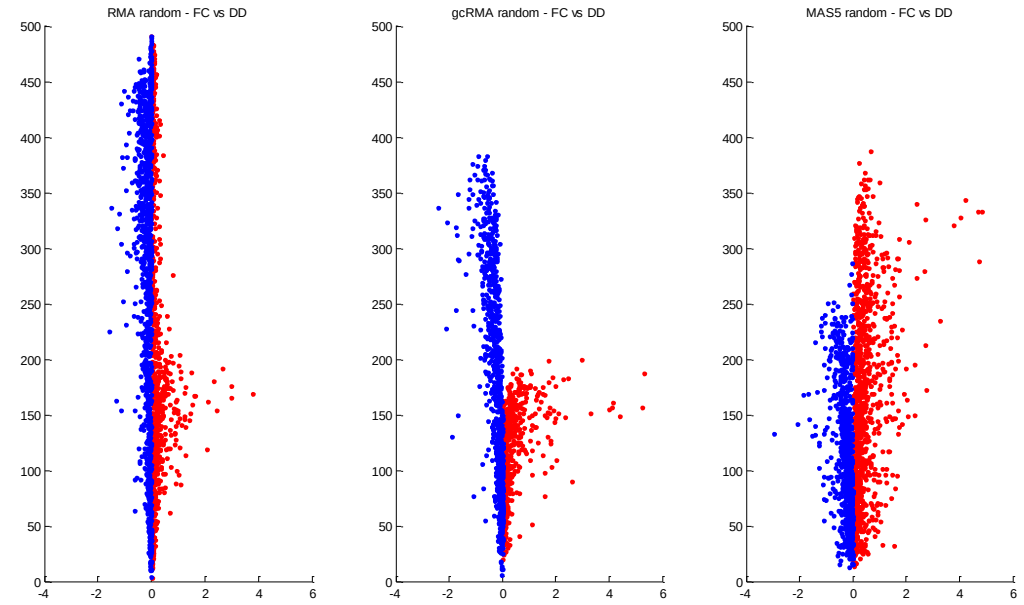


Figure 37. Fold change versus degree distribution for the networks of random data. Red dots represent upregulated genes.

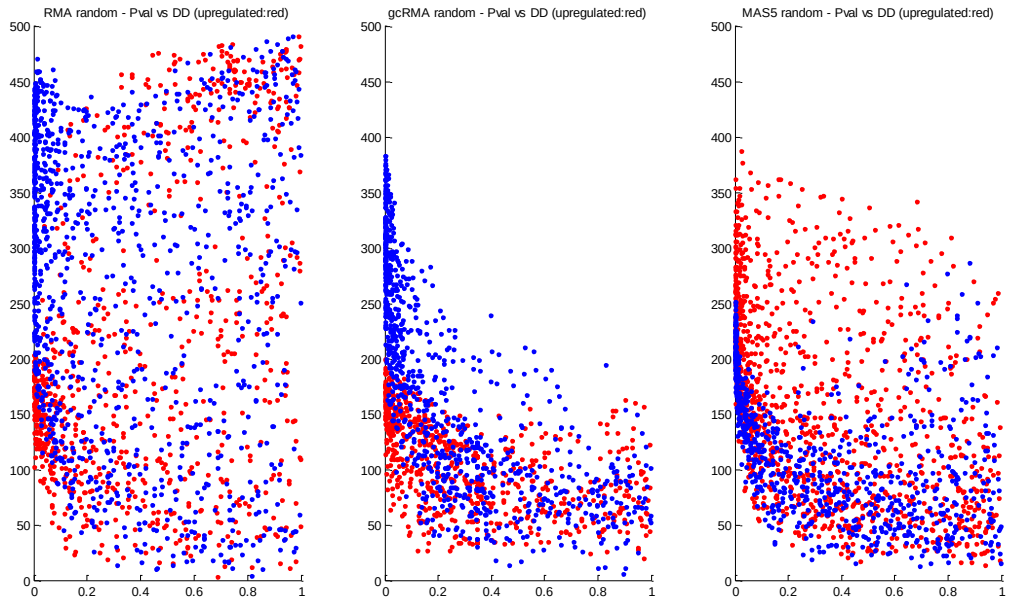


Figure 38. P-value versus degree distribution for the networks of random data. Red dots represent upregulated genes.

To compare different networks of random data among preprocessing methods, scatter plots were drawn (Figures 39-41). Accordingly, there was no significant correlation among any of the network pairs for none of the network topology measures (Table 14).

Table 14. Spearman correlation rho values of pairwise comparison of each network topology measure matrix for random data.

	Clustering coefficient		Betweenness centrality		Degree distributon	
	gcRMA	**MAS5**	**gcRMA**	**MAS5**	**gcRMA**	**MAS5**
RMA	0.10	0.13	0.29	0.17	0.11	0.14
gcRMA		0.08		0.13		0.12

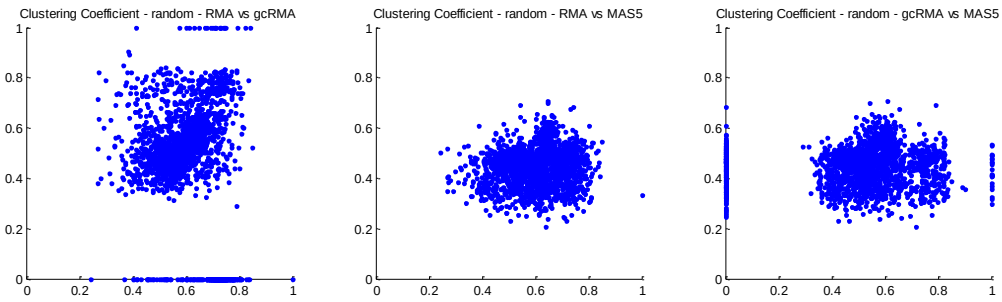


Figure 39. Scatter plots of clustering coefficient values for each network pair or random data. Red dots represent upregulated genes.

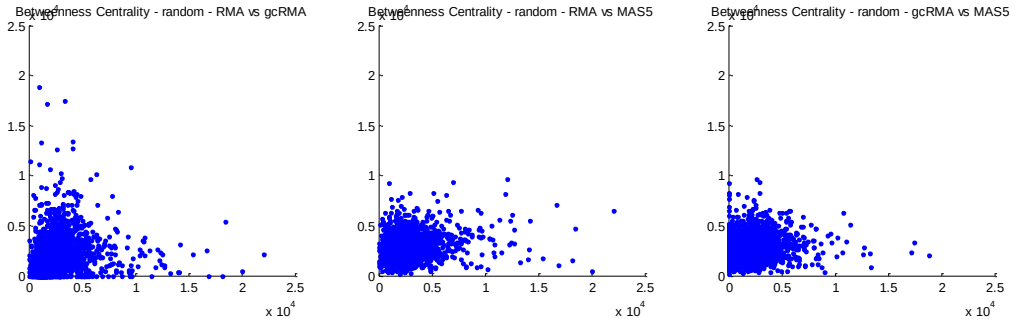


Figure 40. Scatter plots of betweenness centrality values for each network pair or random data. Red dots represent upregulated genes.

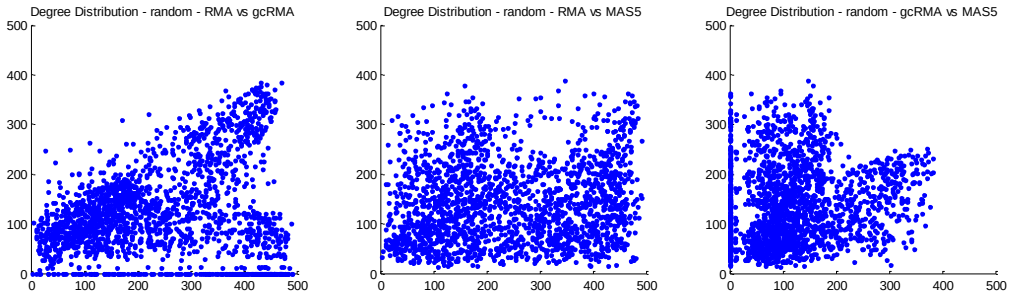


Figure 41. Scatter plots of degree distribution values for each network pair or random data. Red dots represent upregulated genes.

Lastly, to allow a better visual comparison of networks of the intersection and the random data, fold change versus network topology measures were plotted below (Figures 42-44). To conclude, network topology measures that were calculated from the networks of intersection data for each preprocessing method was significantly different from random counterparts. Cellular regulation in response to a condition gives a certain structure to the gene regulatory network and differentiates the network from random data. In addition, random data is highly sensitive to different preprocessing methods so that the similarities between the distributions of network topology measures are highly reduced compared to the network measures of intersection data. Lastly, although network topology measures from networks of different preprocessing methods exhibited similar characteristics compared to the random, there were still differences in terms of values and the distribution of upregulated and downregulated genes.

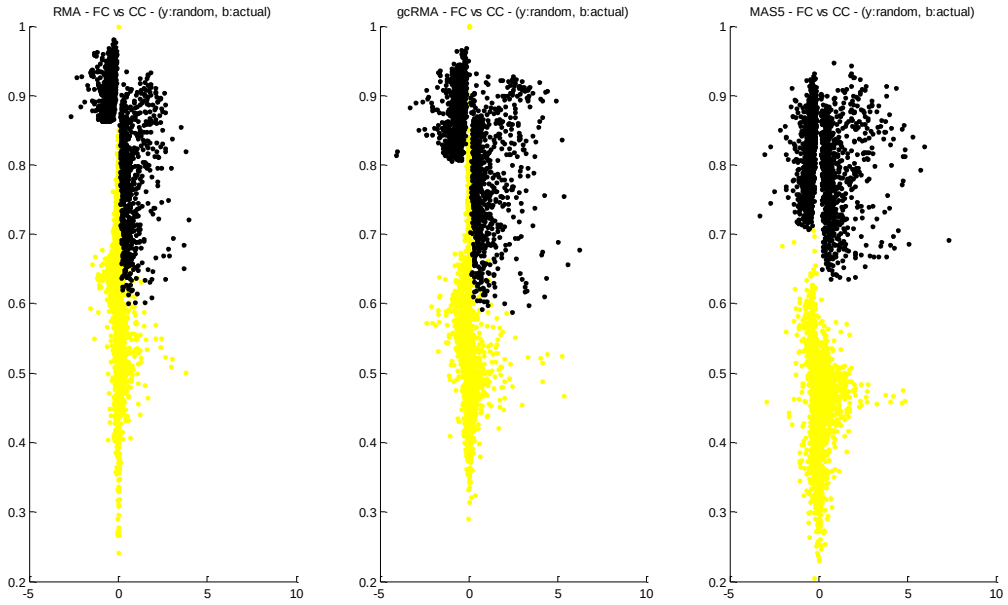


Figure 42. Fold change versus clustering coefficient for both intersection and random networks. Network topology measures from random data is plotted in yellow.

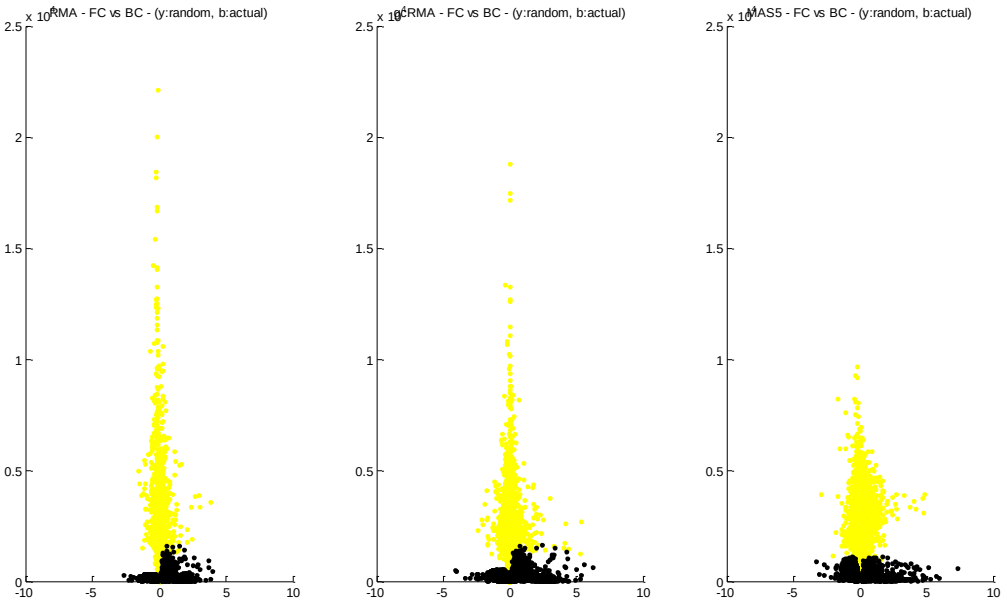


Figure 43. Fold change versus betweenness centrality for both intersection and random networks. Network topology measures from random data is plotted in yellow.

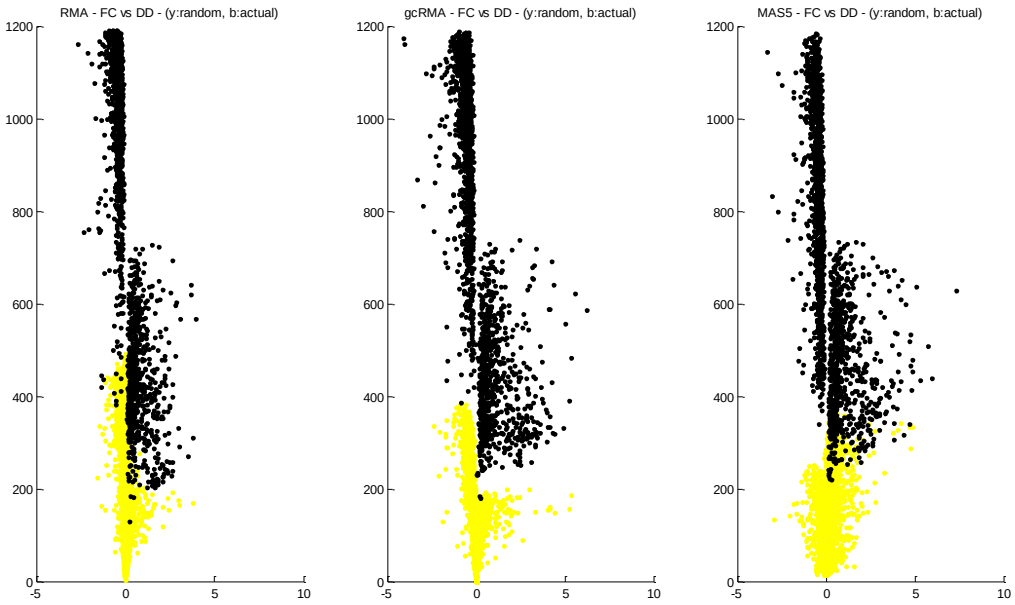


Figure 44. Fold change versus degree distribution for both intersection and random networks. Network topology measures from random data is plotted in yellow.

4.7. Properties of conserved genes in terms of aforementioned network topology measures

Although there were significant differences between the network structures of differently preprocessed data, some genes exhibited similar properties across the networks of RMA, gcRMA, and MAS5 data. To find those genes, the code in APPENDIX F was utilized. Top 20% least-changed genes were identified for both union and intersection data. There are 81 genes in union-based gene list and there are 91 genes in intersection-based gene list (Tables 15, 16). In addition, there are 31 genes that are common for both of the gene lists.

Since the intersection data generated the most variation among the preprocessing methods, 91 genes with the least variation were analyzed in more detail. Table 15 summarizes the network parameters of these 91 genes. These differentially expressed probesets had on average 377, 359, and 375 k-neighbors for RMA, gcRMA, and MAS5 datasets, respectively. Also, mean clustering coefficient values were 0.84, 0.83, and 0.84. Lastly, average betweenness centrality values were 117.4, 118.8, and 114.48.

There were 81 least variant genes in union data according to our criteria (Table 16). Average k-neighbor values for RMA, gcRMA, and MAS5 datasets were 784, 698, and 908. Mean betweenness centrality values were 2484, 2713, and 2527. Lastly, average clustering coefficient values for each dataset were 0.69, 0.66, and 0.67.

Functional enrichment analysis for those genes was conducted using FatiGO feature of Babelomics v3.2 . The top five categories for intersection least variant gene list in GO Biological Process were: 'cellular metabolic process', primary metabolic process', macromolecule metabolic process', 'establishment of localization', and 'regulation of biological quality'. For GO Molecular Function, top five terms were: 'ion binding', 'oxidoreductase activity', enzyme inhibitor activity', 'transferase activity', and 'protein binding'. Lastly, KEGG networks having more than one gene were: 'Tyrosine metabolism' and 'Glycolysis/Gluconeogenesis'.

The top five categories for union least variant gene list in GO Biological Process were: 'establishment of localization', 'cellular metabolic process', 'primary metabolic process', macromolecule metabolic process', and 'regulation of biological process'. For GO Molecular Function, top five terms were: 'ion binding', 'transferase activity', oxidoreductase activity', 'protein binding', and 'nucleotide binding'. Lastly, there were no KEGG networks having at least two genes; there are 11 one-gene networks.

Table 15. Least-variant genes in terms of network topology measures for intersection dataset

Intersection list	RMA					gcRMA					MAS5				
Probe Set ID	**fold change**	**degree**	**clust. coef.**	**betw. cent.**	**p-value**	**fold change**	**degree**	**clust. coef.**	**betw. cent.**	**p-value**	**fold change**	**degree**	**clust. coef.**	**betw. cent.**	**p-value**
Dr.10083.1.S1_at	1.4422	418	0.8320	128.9247	0.0100	2.3589	373	0.8440	99.7848	0.0108	2.7602	365	0.8672	74.8006	0.0130
Dr.10650.1.A1_at	1.6043	355	0.8584	80.9883	0.0439	2.4725	331	0.8584	73.5246	0.0313	2.0586	336	0.8542	78.4487	0.0219
Dr.11033.1.S1_at	0.8783	335	0.8928	50.5002	0.0250	1.2945	300	0.8893	45.8628	0.0464	1.3483	343	0.8824	56.9227	0.0199
Dr.11306.1.S1_at	0.5902	383	0.8624	85.8448	0.0160	0.9144	391	0.8361	114.4221	0.0106	0.6392	410	0.8401	109.7026	0.0127
Dr.11427.1.S1_at	0.5496	507	0.7502	312.9532	0.0044	0.7787	478	0.7336	328.7310	0.0040	0.5910	493	0.7417	305.3502	0.0058
Dr.118.1.S1_at	0.1463	454	0.7640	242.5115	0.0185	0.2628	402	0.7418	245.2834	0.0114	0.2720	411	0.7740	197.7830	0.0193
Dr.1202.1.S1_at	2.0657	352	0.8657	73.2498	0.0089	2.6113	320	0.8685	64.9676	0.0153	1.9281	355	0.8756	63.9683	0.0139
Dr.12107.1.A1_at	0.6074	370	0.8425	98.4312	0.0292	0.8037	338	0.8267	104.6938	0.0114	0.8481	323	0.8081	104.2321	0.0165
Dr.12439.1.S1_at	0.1769	316	0.8182	92.6836	0.0212	0.2132	272	0.8265	70.1738	0.0429	0.1818	312	0.8601	59.2823	0.0391
Dr.1246.1.S1_at	1.4146	389	0.8476	98.7830	0.0122	1.9904	373	0.8310	112.2276	0.0136	1.4807	407	0.8230	130.0214	0.0095
Dr.12525.1.A1_at	0.1592	350	0.8869	56.7750	0.0347	0.1277	292	0.8870	43.1212	0.0301	0.2030	352	0.8597	70.1141	0.0300
Dr.12584.1.S1_at	1.2891	370	0.8419	100.6427	0.0457	1.8685	317	0.8638	64.6751	0.0404	1.5839	325	0.8601	70.3888	0.0249
Dr.12596.1.S1_at	0.5632	426	0.8384	122.8714	0.0110	0.8586	406	0.8194	142.1929	0.0050	0.6418	443	0.8235	145.6160	0.0039
Dr.12602.1.S1_at	1.2793	305	0.9080	37.0254	0.0419	1.7834	313	0.8801	52.8932	0.0407	1.3334	264	0.9078	26.5220	0.0489
Dr.1307.1.S1_at	0.4484	288	0.8976	39.5578	0.0330	0.4617	309	0.8696	59.8006	0.0468	0.9365	313	0.8777	54.1560	0.0289
Dr.13466.1.A1_at	1.1918	294	0.9184	29.2498	0.0346	1.9883	299	0.8942	42.7663	0.0278	2.0138	289	0.9269	24.4773	0.0350
Dr.13681.1.S1_at	2.6589	426	0.8237	136.2818	0.0036	3.6952	395	0.8122	137.6355	0.0047	2.8537	429	0.8183	142.3998	0.0042
Dr.1383.1.S1_at	0.2392	538	0.7462	354.6132	0.0012	0.2268	513	0.7306	384.6579	0.0003	0.1943	556	0.7277	414.0584	0.0013
Dr.14073.1.A1_at	0.4271	332	0.8537	90.9098	0.0323	0.5541	353	0.8456	90.6406	0.0271	0.5240	384	0.8454	98.7177	0.0126

Dr.14502.1.S1_at	0.9319	311	0.8658	63.2195	0.0241	1.6707	341	0.8451	89.9015	0.0280	1.2464	356	0.8297	102.8151	0.0262
Dr.15161.1.S1_at	1.5669	363	0.8757	68.4049	0.0168	2.6125	344	0.8751	64.0300	0.0149	2.6788	308	0.9132	32.8938	0.0255
Dr.15343.1.A1_at	0.3506	356	0.8663	74.1638	0.0155	0.4996	325	0.8779	56.6606	0.0121	0.4346	344	0.8364	85.7960	0.0133
Dr.15373.1.A1_at	0.3034	314	0.8278	92.1330	0.0177	0.2980	290	0.8526	64.6102	0.0399	0.7937	331	0.8464	73.3426	0.0175
Dr.1565.1.S1_at	0.7434	383	0.8480	97.1109	0.0070	1.1047	344	0.8592	76.0591	0.0154	1.1919	390	0.8331	112.4682	0.0102
Dr.1566.1.S1_at	0.3075	371	0.8265	116.1370	0.0398	0.4646	373	0.8218	123.4658	0.0159	0.6343	332	0.8293	94.8577	0.0236
Dr.15949.2.S1_a_at	0.2037	303	0.8276	81.8274	0.0306	0.3267	325	0.8031	109.7830	0.0221	0.2344	309	0.8372	72.3474	0.0416
Dr.1605.1.S1_at	2.4660	379	0.8311	112.7778	0.0059	3.1242	354	0.8221	114.0626	0.0087	2.6294	371	0.8460	92.0051	0.0105
Dr.16720.1.A1_at	1.7378	461	0.7925	209.9562	0.0094	2.4658	440	0.7688	235.9172	0.0053	2.0307	440	0.7899	191.5601	0.0038
Dr.16820.1.S1_at	0.3115	417	0.8508	104.9260	0.0173	0.3915	422	0.8154	151.4586	0.0055	0.4152	452	0.8237	150.7011	0.0043
Dr.169.1.S1_at	0.1534	502	0.7820	243.2909	0.0046	0.1682	447	0.7933	194.4199	0.0056	0.2777	473	0.7858	206.8389	0.0065
Dr.17437.1.S1_at	2.2139	446	0.7965	187.4608	0.0025	3.2136	443	0.7634	236.9798	0.0026	2.6943	482	0.7777	227.9354	0.0025
Dr.17450.2.S1_at	0.4642	286	0.8230	74.3203	0.0277	0.6913	263	0.8410	58.1222	0.0402	0.5555	321	0.8631	59.4253	0.0322
Dr.17459.1.S1_a_at	1.6692	406	0.8345	120.5968	0.0063	2.9326	408	0.7966	169.4862	0.0056	2.7816	437	0.8089	157.4794	0.0054
Dr.17693.1.A1_at	2.0152	357	0.8618	75.7644	0.0093	2.8371	351	0.8369	99.3509	0.0088	2.4861	320	0.8791	54.5175	0.0223
Dr.18175.2.S1_at	0.4555	379	0.7983	166.5328	0.0123	0.6748	387	0.8215	126.7970	0.0137	0.6363	389	0.8427	100.9352	0.0193
Dr.18429.1.A1_at	0.9574	477	0.7813	227.3856	0.0016	1.6394	433	0.7878	190.6312	0.0033	1.5632	458	0.7935	197.2216	0.0028
Dr.18433.1.A1_at	0.4493	406	0.8047	173.2363	0.0119	0.6946	405	0.8256	129.0342	0.0125	0.6512	435	0.8287	133.7239	0.0067
Dr.18473.1.A1_at	1.4449	389	0.8576	89.2140	0.0114	2.2828	402	0.8161	136.4945	0.0050	1.5779	387	0.8424	104.4613	0.0086
Dr.1889.1.S1_at	1.3437	298	0.8905	43.7825	0.0223	1.5946	285	0.8759	51.2847	0.0238	1.4324	302	0.9059	35.3029	0.0276
Dr.1889.2.A1_a_at	1.4612	315	0.8767	54.9238	0.0141	1.8265	319	0.8548	74.9655	0.0134	1.1596	318	0.8714	58.2424	0.0229
Dr.1909.1.S1_at	1.4624	436	0.8046	171.1042	0.0065	2.3807	432	0.7787	206.3316	0.0057	1.7597	454	0.7818	210.3173	0.0050
Dr.19224.1.S1_at	2.2057	487	0.7768	238.8845	0.0012	2.8408	464	0.7609	251.5596	0.0015	2.3043	468	0.7906	202.4760	0.0018
Dr.20125.1.A1_at	1.9969	353	0.8717	68.6326	0.0299	2.7756	353	0.8571	82.1149	0.0199	2.8511	380	0.8388	103.9149	0.0137
Dr.20125.1.A1_s_at	2.1182	368	0.8608	81.3666	0.0283	3.0631	375	0.8296	115.9568	0.0189	2.9974	394	0.8258	125.4568	0.0090
Dr.20198.2.S1_x_at	0.4974	362	0.8763	67.2388	0.0333	0.8082	339	0.8662	68.7801	0.0177	0.6738	346	0.8570	73.4149	0.0142

Dr.20270.1.S1_at	2.6310	400	0.8198	139.8986	0.0032	3.6010	341	0.8415	92.0049	0.0101	2.9382	386	0.8487	94.2655	0.0098
Dr.20291.1.A1_at	2.2522	424	0.8119	154.0834	0.0036	3.3508	402	0.7956	167.0332	0.0044	2.7028	432	0.8097	156.1078	0.0040
Dr.2059.1.A1_at	1.2208	396	0.8495	100.9117	0.0140	2.0363	371	0.8485	93.9231	0.0093	2.5424	428	0.8186	140.7996	0.0046
Dr.20610.1.S1_at	1.4995	456	0.7990	191.4333	0.0061	2.1309	448	0.7713	228.0377	0.0037	1.8052	456	0.7763	222.5600	0.0051
Dr.21064.1.S1_at	1.7548	326	0.8768	58.5721	0.0126	2.3248	313	0.8681	64.0348	0.0160	1.9659	322	0.8991	43.0623	0.0224
Dr.21879.1.A1_at	0.4181	378	0.8271	118.0150	0.0482	0.5297	342	0.8533	81.5824	0.0213	0.5184	358	0.8257	111.3994	0.0157
Dr.22139.1.A1_at	1.6390	393	0.8166	145.3264	0.0032	2.5416	354	0.8304	106.8621	0.0079	1.8801	352	0.8568	79.6681	0.0171
Dr.2426.1.S1_at	1.5811	287	0.8939	40.4865	0.0299	2.3374	254	0.9045	31.2350	0.0412	2.0914	305	0.8978	41.5365	0.0309
Dr.24311.1.S1_at	1.1960	542	0.7070	496.5489	0.0001	2.2782	555	0.6866	550.0544	0.0002	1.6379	595	0.7002	550.6426	0.0010
Dr.2452.1.A1_at	1.6102	457	0.7960	193.3901	0.0046	2.8619	416	0.7918	175.1884	0.0037	1.9932	432	0.8029	171.8218	0.0044
Dr.2452.2.A1_x_at	2.2113	440	0.8123	157.5769	0.0043	3.2116	386	0.8170	129.3305	0.0056	2.5723	437	0.8084	159.1714	0.0036
Dr.2528.2.S1_at	0.1463	387	0.7684	207.3199	0.0095	0.2306	405	0.7963	173.4559	0.0126	0.3721	351	0.7851	137.7906	0.0240
Dr.25822.1.S1_at	0.2835	252	0.8426	54.6969	0.0299	0.3403	242	0.8577	43.9518	0.0486	0.2689	267	0.8673	42.0610	0.0456
Dr.25893.1.A1_at	1.5359	356	0.8630	77.6803	0.0373	2.3880	355	0.8516	87.7165	0.0167	2.1272	353	0.8608	78.1277	0.0157
Dr.2596.1.S1_a_at	0.9097	353	0.8765	63.6742	0.0191	1.1604	320	0.8769	57.2367	0.0198	0.8835	321	0.8974	44.7333	0.0262
Dr.26328.1.A1_at	0.7039	397	0.8214	129.6359	0.0292	1.0736	358	0.8173	117.8454	0.0195	0.8463	349	0.8297	100.0420	0.0165
Dr.26428.1.A1_at	0.2560	414	0.7950	203.2180	0.0184	0.3529	457	0.7864	213.5322	0.0049	0.3159	468	0.8070	181.6931	0.0048
Dr.2890.1.A1_at	0.3275	418	0.8171	147.4048	0.0066	0.3795	377	0.8337	111.2460	0.0119	0.5614	371	0.8240	112.4227	0.0147
Dr.2960.1.A1_at	1.7664	411	0.8317	122.3043	0.0045	2.8388	358	0.8463	91.6383	0.0090	2.2778	397	0.8386	107.5812	0.0081
Dr.3004.1.A1_at	1.5983	307	0.8878	47.0386	0.0268	2.4313	268	0.9009	33.2461	0.0424	1.9405	297	0.8871	45.6818	0.0340
Dr.3025.3.S1_at	1.6400	431	0.8247	139.1880	0.0054	2.9047	380	0.8383	107.0443	0.0090	2.1874	432	0.8148	150.2986	0.0037
Dr.3374.2.S1_at	0.3753	266	0.8170	88.0121	0.0331	0.4624	305	0.8407	87.3886	0.0341	0.3386	318	0.8491	71.3644	0.0372
Dr.3529.1.S1_at	0.9687	331	0.8741	61.6824	0.0143	1.7289	334	0.8664	68.4174	0.0195	1.0901	284	0.9074	33.4027	0.0457
Dr.3613.1.S1_at	1.6917	314	0.9077	38.2392	0.0243	2.4522	284	0.9206	27.2448	0.0326	1.7892	289	0.9293	23.9456	0.0405
Dr.382.2.S1_at	0.9810	426	0.8224	146.0384	0.0130	1.2618	390	0.8223	129.3607	0.0121	1.1171	437	0.7951	184.7048	0.0043
Dr.4111.1.S1_at	1.7983	383	0.8475	96.0250	0.0058	2.7542	330	0.8607	73.3332	0.0117	2.2144	375	0.8547	85.5516	0.0104

Dr.4744.1.S1_a_at	1.7621	352	0.8675	71.2388	0.0333	2.8378	354	0.8412	95.4053	0.0278	2.1339	347	0.8523	81.4562	0.0211
Dr.4907.1.S1_at	2.4190	322	0.8810	55.4930	0.0158	3.3166	291	0.8932	43.3212	0.0263	2.5360	351	0.8754	63.2060	0.0158
Dr.4961.1.A1_at	1.6086	370	0.8358	106.9488	0.0076	2.6971	373	0.8124	133.5757	0.0076	2.0121	369	0.8336	105.4383	0.0179
Dr.4968.1.A1_at	0.3566	446	0.8065	176.2252	0.0208	0.5649	418	0.7921	187.1157	0.0047	0.3132	459	0.7875	206.6735	0.0035
Dr.4975.1.A1_at	1.0895	501	0.7670	284.2636	0.0091	1.5322	472	0.7505	294.3565	0.0033	1.5222	506	0.7588	295.8125	0.0012
Dr.5462.1.S1_at	2.2073	317	0.8917	45.4384	0.0172	3.0788	257	0.9101	29.6479	0.0370	2.3117	295	0.9229	26.3572	0.0327
Dr.5467.1.A1_at	1.4544	327	0.8925	48.7453	0.0224	2.2720	310	0.8957	43.4191	0.0233	1.8206	310	0.9132	33.5952	0.0294
Dr.5562.1.S1_at	1.8048	308	0.8597	66.8386	0.0207	2.3576	274	0.8593	57.3169	0.0245	1.8471	315	0.8674	59.4095	0.0232
Dr.5674.2.S1_at	1.7958	276	0.9017	33.3377	0.0265	2.1870	260	0.8990	35.2253	0.0312	1.9289	257	0.9267	21.9633	0.0488
Dr.6550.1.A1_at	0.7249	387	0.8551	92.7600	0.0142	1.1912	366	0.8413	101.9364	0.0149	0.9850	395	0.8386	110.0411	0.0087
Dr.6787.1.S1_at	1.3627	405	0.8359	119.2509	0.0119	2.2903	394	0.8241	126.7245	0.0071	1.5177	352	0.8560	80.0484	0.0146
Dr.7171.1.S1_at	0.9196	362	0.8702	72.2950	0.0215	1.4182	368	0.8361	105.8985	0.0148	1.2390	395	0.8326	114.4521	0.0098
Dr.7599.1.A1_at	1.7602	378	0.8592	83.5027	0.0077	2.2994	342	0.8603	75.4826	0.0103	2.0128	358	0.8758	64.0533	0.0118
Dr.7692.1.A1_at	1.2143	402	0.8448	107.3664	0.0153	1.8535	349	0.8621	76.9618	0.0236	1.8199	403	0.8206	135.4845	0.0072
Dr.845.1.A1_at	1.8391	266	0.9336	20.0143	0.0373	2.4863	252	0.9276	20.7175	0.0464	1.8661	268	0.9427	16.0624	0.0469
Dr.848.1.S1_at	1.4282	322	0.9018	42.4586	0.0281	2.0632	325	0.8835	54.6485	0.0207	1.7777	335	0.8882	50.8818	0.0174
Dr.8750.1.A1_at	0.4569	411	0.8221	141.4153	0.0061	0.6682	435	0.7872	198.5422	0.0047	0.5562	437	0.8087	161.2360	0.0054
Dr.8947.1.A1_at	1.1202	423	0.8122	159.9608	0.0206	1.6358	398	0.8071	153.3707	0.0081	1.2976	387	0.8271	121.8439	0.0111
Dr.9025.1.A1_at	1.0317	262	0.9056	32.2742	0.0444	1.4788	247	0.8809	44.2285	0.0433	1.3835	300	0.8904	44.7290	0.0347
Dr.938.1.S1_at	0.4423	442	0.8078	172.5354	0.0140	0.6179	450	0.7674	235.3578	0.0046	0.5001	418	0.7954	167.2451	0.0096

Table 16. Least-variant genes in terms of network topology measures for union dataset

Union list	RMA					gcRMA					MAS5				
Probe Set ID	**fold change**	**degree**	**clust. coef.**	**betw. cent.**	**p-value**	**fold change**	**degree**	**clust. coef.**	**betw. cent.**	**p-value**	**fold change**	**degree**	**clust. coef.**	**betw. cent.**	**p-value**
Dr.10283.1.A1_at	0.5987	1211	0.5403	6058.2121	0.0001	0.7234	1175	0.5028	7089.0780	0.0000	0.5529	1535	0.5500	5703.3336	0.0001
Dr.1064.1.S1_at	0.8697	702	0.7279	1369.0736	0.0224	1.6003	587	0.6951	1634.7103	0.0347	1.2515	834	0.6961	1260.2518	0.0137
Dr.11022.1.A1_at	0.2447	726	0.6665	3418.2778	0.0355	0.3399	685	0.6137	2730.9087	0.0083	0.3093	849	0.6370	3902.4308	0.0147
Dr.11302.1.A1_at	-0.5253	1167	0.7663	2889.3121	0.0022	-0.8441	810	0.7855	2732.6287	0.0507	-0.7422	844	0.7561	1646.0926	0.0121
Dr.11551.1.S1_at	0.5158	612	0.7219	1392.0663	0.0614	0.6814	560	0.7364	1936.3196	0.1125	1.2611	906	0.7666	2240.9364	0.0443
Dr.1201.1.S1_at	-0.1659	925	0.7951	1892.3787	0.0099	-0.1873	702	0.8240	1965.1412	0.0411	-0.2184	787	0.7622	2271.8737	0.0163
Dr.1202.1.S1_at	2.0657	678	0.7261	1003.6984	0.0089	2.6113	533	0.7434	2473.0475	0.0153	1.9281	853	0.6957	1207.1885	0.0139
Dr.12227.1.A1_at	0.7429	655	0.7653	782.2777	0.0452	1.0512	554	0.7236	953.3935	0.0469	1.0193	838	0.7033	1154.2571	0.0141
Dr.12439.6.S1_at	0.2453	1034	0.5633	5766.3740	0.0012	0.3203	1106	0.5313	5209.6086	0.0002	0.2096	1113	0.5949	4936.8284	0.0062
Dr.1246.1.S1_at	1.4146	781	0.6976	1492.6945	0.0122	1.9904	735	0.6465	1816.0787	0.0136	1.4807	971	0.6560	2174.0725	0.0095
Dr.12671.1.S1_at	0.2205	665	0.6731	3842.9678	0.0786	0.4617	412	0.7029	2269.9528	0.0835	0.3551	681	0.6866	2303.1738	0.0346
Dr.13635.1.S1_at	0.2841	451	0.7084	1119.2865	0.0998	0.5511	604	0.7117	1991.6230	0.0869	0.5719	745	0.7611	1936.4097	0.0461
Dr.13681.1.S1_at	2.6589	831	0.6807	1470.3795	0.0036	3.6952	748	0.6438	1720.9211	0.0047	2.8537	1046	0.6412	2401.4048	0.0042
Dr.14502.1.S1_at	0.9319	651	0.6918	1652.5514	0.0241	1.6707	653	0.6421	1905.6200	0.0280	1.2464	913	0.6546	1975.3856	0.0262
Dr.1458.1.S1_at	0.7937	839	0.6400	3774.4634	0.0111	1.1560	925	0.5905	3008.8399	0.0047	1.0413	1201	0.6314	3592.2167	0.0061
Dr.15050.2.S1_at	-0.1663	408	0.7761	3418.0436	0.0475	-0.0968	471	0.7783	3260.2699	0.0889	-0.3325	475	0.7757	1996.4016	0.0473
Dr.15161.1.S1_at	1.5669	734	0.7289	1200.1159	0.0168	2.6125	691	0.6793	1265.0392	0.0149	2.6788	760	0.7174	1101.9825	0.0255
Dr.15373.1.A1_at	0.3034	620	0.6369	2588.6622	0.0177	0.2980	597	0.6461	3503.0321	0.0399	0.7937	895	0.6575	2189.8161	0.0175
Dr.15824.1.S1_at	0.2528	829	0.6468	3876.1568	0.0387	0.3108	638	0.6459	3551.1544	0.0777	0.9388	698	0.6868	2387.5348	0.0326
Dr.15949.2.S1_a_at	0.2037	616	0.6258	2510.6725	0.0306	0.3267	573	0.6319	3500.7494	0.0221	0.2344	916	0.6657	3284.6624	0.0416

Dr.1605.1.S1_at	2.4660	708	0.6921	1169.9543	0.0059	3.1242	568	0.7076	838.7443	0.0087	2.6294	886	0.6742	1425.5731	0.0105
Dr.16687.2.A1_at	0.2474	755	0.7111	5644.3346	0.0406	0.5342	648	0.6657	4687.8877	0.0651	0.2951	626	0.7279	5608.1156	0.0741
Dr.17459.1.S1_a_at	1.6692	813	0.6759	1718.1232	0.0063	2.9326	774	0.6273	1934.0320	0.0056	2.7816	1086	0.6342	2421.5845	0.0054
Dr.1756.1.S1_at	0.1985	808	0.6067	3592.7297	0.0172	0.2652	812	0.5430	4665.5239	0.0283	0.2541	1090	0.6044	3130.7788	0.0092
Dr.18429.1.A1_at	0.9574	940	0.6238	2481.8712	0.0016	1.6394	860	0.5980	2605.2085	0.0033	1.5632	1072	0.6369	2203.2984	0.0028
Dr.1889.1.S1_at	1.3437	566	0.7388	1318.9764	0.0223	1.5946	464	0.7542	638.1520	0.0238	1.4324	698	0.7308	790.4809	0.0276
Dr.1889.2.A1_a_at	1.4612	602	0.7270	1036.0476	0.0141	1.8265	534	0.7195	754.9281	0.0134	1.1596	782	0.6870	1083.2679	0.0229
Dr.19224.1.S1_at	2.2057	936	0.6331	2324.0187	0.0012	2.8408	885	0.5871	2739.0937	0.0015	2.3043	1122	0.6226	2594.6324	0.0018
Dr.20010.3.S1_at	0.1431	679	0.6441	3668.5484	0.0095	0.1725	749	0.6056	3523.6622	0.0061	0.1416	1042	0.6209	4164.5639	0.0186
Dr.20291.1.A1_at	2.2522	819	0.6643	1721.1783	0.0036	3.3508	695	0.6509	1514.1011	0.0044	2.7028	1039	0.6413	1965.4362	0.0040
Dr.2059.1.A1_at	1.2208	814	0.6922	1602.6744	0.0140	2.0363	760	0.6463	1899.2521	0.0093	2.5424	1030	0.6506	1914.2302	0.0046
Dr.21064.1.S1_at	1.7548	636	0.7323	1035.0732	0.0126	2.3248	519	0.7393	2437.9260	0.0160	1.9659	795	0.7045	1366.8609	0.0224
Dr.2117.1.S1_at	1.2772	763	0.7059	3635.9734	0.0391	2.1448	615	0.6648	3347.3567	0.0960	1.8154	720	0.7273	2503.1325	0.0337
Dr.22100.1.A1_at	1.8940	786	0.6348	2658.1522	0.0085	2.9863	812	0.5836	3609.3688	0.0033	1.9541	1024	0.6224	2488.9895	0.0047
Dr.22139.1.A1_at	1.6390	774	0.6450	2484.9647	0.0032	2.5416	582	0.7084	3133.1717	0.0079	1.8801	919	0.6691	1606.8244	0.0171
Dr.22721.1.A1_at	-0.3477	820	0.8084	1915.2492	0.0138	-0.3602	587	0.7935	2179.9547	0.0871	-0.4364	450	0.7688	3488.3973	0.0882
Dr.24220.1.A1_at	0.1351	617	0.6864	2744.2351	0.0772	0.1666	572	0.6293	3819.2617	0.0646	0.1770	907	0.6333	3377.6077	0.0103
Dr.2452.1.A1_at	1.6102	891	0.6548	2520.7213	0.0046	2.8619	790	0.6219	2397.4926	0.0037	1.9932	1003	0.6453	2147.5685	0.0044
Dr.2452.2.A1_a_at	2.0261	860	0.6657	2329.5106	0.0063	3.0575	744	0.6357	2180.5373	0.0063	2.4560	809	0.6994	1138.0723	0.0143
Dr.2452.2.A1_x_at	2.2113	846	0.6761	1591.6557	0.0043	3.2116	721	0.6509	1802.2186	0.0056	2.5723	1028	0.6472	2039.3736	0.0036
Dr.24867.2.A1_at	0.3427	686	0.7472	1493.4852	0.0480	0.3258	490	0.7218	1972.0357	0.1195	1.2124	690	0.7456	802.8607	0.0349
Dr.25607.1.S1_at	0.4249	1084	0.5702	4491.9567	0.0009	0.5836	976	0.5339	6001.4454	0.0021	0.4495	1233	0.5796	5356.3056	0.0075
Dr.25759.2.A1_a_at	0.4686	1190	0.5378	6006.7261	0.0001	0.6632	1076	0.5192	5776.3145	0.0000	0.4880	1317	0.5789	4366.3795	0.0014
Dr.2596.1.S1_a_at	0.9097	704	0.7417	788.2018	0.0191	1.1604	639	0.6910	1319.6866	0.0198	0.8835	759	0.7160	1334.3166	0.0262

Dr.2596.3.A1_at	0.8335	646	0.7669	1016.1392	0.0420	1.1160	580	0.7227	1048.7299	0.0564	0.8958	660	0.7434	981.4599	0.0452
Dr.26025.1.A1_at	0.1809	555	0.6507	1671.4191	0.0611	0.2445	501	0.6478	1337.9574	0.0549	0.2145	808	0.6834	1881.5532	0.0395
Dr.26266.1.A1_at	0.5621	716	0.7276	2122.6997	0.0374	0.6742	578	0.6843	2793.0832	0.1160	0.7498	604	0.6910	1913.6797	0.0878
Dr.26321.1.S1_at	-0.1226	735	0.8080	3467.7072	0.0467	-0.0747	637	0.7865	4511.3919	0.2271	0.1760	725	0.7441	4692.7345	0.1029
Dr.26343.1.A1_at	0.4737	975	0.5941	5056.0425	0.0023	0.6392	1031	0.5550	4241.0024	0.0005	0.5581	1316	0.5914	5304.7879	0.0010
Dr.3004.1.A1_at	1.5983	641	0.7260	1029.8433	0.0268	2.4313	496	0.7253	775.2294	0.0424	1.9405	684	0.7159	1584.6930	0.0340
Dr.3025.2.S1_at	1.1051	790	0.6966	1291.9058	0.0077	1.7346	715	0.6641	1443.9563	0.0107	1.9000	843	0.6910	1679.3797	0.0196
Dr.3025.3.S1_at	1.6400	866	0.6685	2023.6693	0.0054	2.9047	786	0.6303	2350.2076	0.0090	2.1874	1036	0.6457	2046.1160	0.0037
Dr.3436.1.A1_at	0.9481	886	0.6511	3561.2310	0.0129	1.4025	598	0.6720	3510.0171	0.1018	1.6881	787	0.6886	3475.9938	0.0170
Dr.4002.1.A1_at	1.4124	813	0.6697	2472.1012	0.0116	2.2227	657	0.6695	1620.6445	0.0127	1.6863	771	0.6955	1476.9336	0.0172
Dr.4094.1.S1_at	0.2353	636	0.7261	1255.1763	0.0300	0.0995	482	0.7447	2420.4900	0.1771	0.6340	711	0.6821	1138.0484	0.0313
Dr.4111.1.S1_at	1.7983	739	0.7014	1173.3239	0.0058	2.7542	560	0.7237	704.6941	0.0117	2.2144	923	0.6772	1325.3613	0.0104
Dr.4186.1.S1_at	0.3994	879	0.6312	4430.7344	0.0094	0.4702	956	0.5825	3368.2092	0.0035	0.4473	1174	0.6297	3703.8633	0.0038
Dr.4723.1.A1_x_at	-0.4831	1164	0.8163	1213.7157	0.0090	-0.4878	910	0.8047	1449.0434	0.0232	-0.5842	808	0.7615	1898.5578	0.0156
Dr.4867.1.A1_at	1.9146	1057	0.5719	4393.0340	0.0008	2.5950	1021	0.5339	4659.1799	0.0004	2.1851	1372	0.5785	4067.3191	0.0006
Dr.523.1.A1_at	0.4412	751	0.6718	1912.1156	0.0096	0.6330	626	0.6368	3345.2207	0.0048	0.5255	937	0.6347	2042.8686	0.0089
Dr.5462.1.S1_at	2.2073	620	0.7581	699.4417	0.0172	3.0788	429	0.7893	1833.2971	0.0370	2.3117	714	0.7375	945.7710	0.0327
Dr.5562.1.S1_at	1.8048	579	0.7175	1328.3704	0.0207	2.3576	457	0.7266	812.8500	0.0245	1.8471	718	0.7029	1030.5393	0.0232
Dr.5674.2.S1_at	1.7958	512	0.7704	637.9079	0.0265	2.1870	424	0.7823	2055.6611	0.0312	1.9289	600	0.7582	630.1531	0.0488
Dr.6007.1.S1_at	1.2442	875	0.6586	3196.8687	0.0103	1.9371	831	0.5983	2648.1591	0.0071	1.6238	1040	0.6400	2425.6604	0.0053
Dr.6349.1.A1_at	0.8398	833	0.6555	2157.9793	0.0022	1.4110	783	0.6171	2099.5578	0.0027	1.9916	1008	0.6535	1980.6415	0.0073
Dr.6550.1.A1_at	0.7249	795	0.6967	1531.1854	0.0142	1.1912	725	0.6443	1669.9272	0.0149	0.9850	986	0.6476	2882.2804	0.0087
Dr.6787.1.S1_at	1.3627	839	0.6734	2115.4994	0.0119	2.2903	805	0.6235	2477.7169	0.0071	1.5177	871	0.6810	1230.1938	0.0146
Dr.6807.1.S1_at	0.1051	644	0.7042	2882.1534	0.0837	0.2465	693	0.6655	1943.7223	0.0134	0.2423	1011	0.6761	3289.8490	0.0228
Dr.701.1.S1_at	0.1127	564	0.7049	7964.8691	0.1839	0.1562	622	0.6467	9335.6527	0.0940	0.2222	799	0.7077	8074.2616	0.0444
Dr.728.4.S1_at	1.6042	426	0.7986	677.5884	0.0470	2.0273	362	0.8022	1622.0957	0.0580	1.5748	540	0.7828	1371.7265	0.0774

Dr.7599.1.A1_at	1.7602	726	0.7242	893.5280	0.0077	2.2994	624	0.6917	1069.3522	0.0103	2.0128	857	0.6998	1191.8177	0.0118
Dr.7722.1.A1_at	1.3694	873	0.6440	2212.2793	0.0029	2.1334	835	0.5994	2463.1252	0.0025	1.7676	1189	0.6123	3050.7932	0.0044
Dr.845.1.A1_at	1.8391	541	0.8104	451.6694	0.0373	2.4863	450	0.7767	1916.4866	0.0464	1.8661	643	0.7600	702.4166	0.0469
Dr.8497.1.A1_at	1.1335	593	0.7914	657.9382	0.0577	1.8064	527	0.7406	2096.2970	0.0608	1.5631	739	0.7287	1148.6928	0.0364
Dr.8516.1.S1_at	1.7795	704	0.7261	959.0614	0.0105	2.5894	573	0.7174	776.5566	0.0140	2.0569	916	0.6757	1359.3390	0.0095
Dr.8587.1.A1_at	1.4329	1228	0.5392	6437.8991	0.0002	1.8579	1127	0.5080	7866.1942	0.0001	1.4803	1481	0.5422	6756.7905	0.0001
Dr.8587.1.A2_at	1.4169	1210	0.5430	6079.5360	0.0002	1.8390	1080	0.5197	6559.9418	0.0001	1.3805	1446	0.5506	6765.5736	0.0002
Dr.8750.1.A1_at	0.4569	840	0.6228	3446.3214	0.0061	0.6682	859	0.5952	4485.2267	0.0047	0.5562	1191	0.6376	3808.6929	0.0054
Dr.885.1.S1_at	-0.4508	1069	0.8077	1909.1040	0.0129	-0.5606	872	0.7708	1946.3783	0.0319	-0.5674	801	0.7491	2310.2945	0.0195
Dr.994.3.S1_at	-0.5050	1449	0.7673	3037.9088	0.0008	-0.5969	1339	0.7416	2217.8731	0.0019	-0.6672	1113	0.7048	2951.4539	0.0010
DrAffx.2.103.S1_at	0.5964	534	0.7222	2347.1801	0.1095	0.5300	416	0.7365	2055.2755	0.3026	1.3252	749	0.7048	2601.4420	0.0451

CHAPTER 5: DISCUSSION

Microarray data give quantitative yet relative information about a gene's expression (Irizarry, Hobbs, *et al.* 2003; Millenaar *et al.* 2006). Thus, expression values in each array should be standardized when comparing a batch of arrays. In addition, microarray experiment is prone to technical errors which do not have a biological source (Hill *et al.* 2001; Bilban *et al.* 2002; Grewal *et al.* 2007; Kreil *et al.* 2005; Zakharkin *et al.* 2005). Separation of technical noise from biological variation poses a challenge for the analysis of microarray data (Quackenbush 2002; Boes *et al.* 2005; Giles *et al.* 2003). To overcome these problems, different preprocessing algorithms (RMA, gcRMA, MAS5 etc.) have been developed. Each algorithm aims to filter a typical noise by modifying the raw data. In this study, RMA, gcRMA, and MAS5 were used for a comparison of the preprocessing methods in terms of data distribution, differential expression lists, and network parameters using the Zebrafish GeneChip array. This study is the first to compare different preprocessing methods in an exemplary zebrafish gene expression data (i.e., GSE4989). Furthermore, it proposes novel methods of exploratory analysis of network parameters to understand the nature of correlations in up- and down-regulated probesets in any two-group expression data.

Since there is no reference dataset for Affymetrix Zebrafish GeneChip arrays, the dataset that was used for the present study has been examined in terms of array quality based on selected R packages and BRB-ArrayTools (APPENDIX I). Accordingly, the profiles of RNA degradation plots for normoxia and hypoxia were similar. The percent presence ratio of the probesets ranged between 64.5 and 72.1 percent and the background level of the arrays were low and comparable; these indicated that arrays had high quality hybridization (APPENDIX I). According to affyPLM package of array quality control, images of some arrays exhibited blotches (data not shown). Although NUSE plots and statistics indicated consistent probeset-wise variability across arrays, two arrays, namely GSM112800 and GSM112806 had relatively greater variability (APPENDIX I). However, since the GSE4989 dataset had considerably high number of replicates for each group (n=5) and our methodology depended on gene network comparisons of significantly differentially

expressed genes ($p<0.05$), the effects of array quality differences were considered minimal. In addition, the scatterplot comparisons of the full dataset and the reduced dataset excluding GSM112800 and GSM112806 revealed that data points along the diagonal line were distributed similarly between plots (APPENDIX I). Future studies might include testing of the effects of array quality on the network structure using a leave-one-out crossvalidation approach.

Comparative correlation structure and DEG analysis

RMA method considers only the perfect match (PM) probes of the Affymetrix GeneChip array and performs quantile normalization for a better standardization of arrays (Irizarry, Bolstad, *et al.* 2003). gcRMA is a modified version of RMA so that takes into account the GC content of each probe for removing non-biological noise where the method MAS5 is a scaling approach considering perfect match probes after the effect of mismatch probe intensity is removed (Binder *et al.* 2010). Ploner *et al.* (2005) suggested that the differences in the background correction step might result in the observed statistical differences between the preprocessed data. Similarly, all three methods used in this study exhibited characteristic differences in terms of the parameters examined, which partly may be due to the background adjustment differences.

It is essential to understand the change in the nature of data before making any comparisons and/or performing a statistical test. The differences observed among the preprocessed datasets were striking and showed that preprocessing might drastically alter the data distribution and location of median (Figures 6, 7, and 8). Since statistical tests for identification of differentially expressed genes, for example t-test, may make use of the mean and standard error, alterations in those values dramatically change list of significant genes (Ploner *et al.* 2005). Furthermore, t-test requires assumptions on distribution and variance of the data to be compared (Murie *et al.* 2009). Since MAS5 has a scaling approach, MAS5-preprocessed data might be affected by the upwards or downwards shift of medians in arrays (Binder *et al.* 2010). Existence of shifts or alterations still after the preprocessing would affect the performance of the test (Freeman *et al.* 2007). Compared to MAS5 preprocessed data, RMA and gcRMA standardize the data much more efficiently and yield reliable results when identifying

differentially expressed genes; this is also stressed in the literature (Bolstad *et al.* 2003; Autio *et al.* 2009; Lim *et al.* 2007). Although RMA and gcRMA more efficiently standardize the general characteristics of arrays, they alter the data dramatically as shown in Figures 1-4. Thus, for the statistical analysis based on the correlation of data like clustering or classification, MAS5 has been favored (Lim *et al.* 2007). On the other hand, realtime qRT-PCR confirmation of microarray analyses suggested that MAS5, gcRMA, and dChip might be favorable for medium and high-intensity genes (Qin *et al.* 2006). Similarly transcripts found changing by MAS5 but not gcRMA and RMA could be confirmed by qRT-PCR (Pepper *et al.* 2007). On the other hand, RMA produced the most reproducible results with highest correlation with the qRTPCR in other studies (Millenaar *et al.* 2006).

Other methods of comparing preprocessing algorithms include study of spike-in datasets, measurement of FDR and randomization (Lim *et al.* 2007; Shedden *et al.* 2005; Vardhanabhuti *et al.* 2006). For example, spike-in dataset analysis suggested that a combination of gcRMA with Cyber-T or SAM performed the best for identification of differentially expressed genes (Vardhanabhuti *et al.* 2006).

Randomization of actual expression network was also used to test across different preprocessing. It was found that background adjustment of gcRMA was problematic, truncation of expression resulted in flaws in analysis since adjusted gcRMA was able to fix these problems. Accordingly the authors suggested that normalization affected correlation structure and their results favored MAS5 (Lim *et al.* 2007).

As a result, different studies and analytic approaches favor different preprocessing algorithms thus it is important to apply an appropriate algorithm considering the type of the dataset and its correlation structure. Our findings further emphasize that the correlation structure of the datasets is affected by preprocessing methodology thus one should be aware of its extent for the dataset in consideration before generation of differentially expressed gene lists.

Identification of differentially expressed genes was performed using a two-sample t-test in this study. Although this is a parameteric test and a raw p-value threshold of 0.05 has been used with higher false discovery rate, this is justifyable since it allows for selection of a large number of probesets in the intersection dataset where all preprocessing algorithms resulted in a significant call. According to our results MAS5

provided the highest number of genes in contrast to gcRMA (Table 1). When the intersection of gene lists was compared, RMA and gcRMA exhibited a greater amount of overlap. This was due to the nature of the mentioned preprocessing methods where the only difference was in background correction approach (Binder *et al.* 2010). Also in various studies, it is shown that differentially expressed gene lists from RMA and gcRMA preprocessed data has higher similarity compared to MAS5 (Binder *et al.* 2010; Lim *et al.* 2007; Freeman *et al.* 2007).

In the literature, there are many studies demonstrating that the identity and number of differentially expressed genes can be influenced by the preprocessing method chosen. For example, genes altered in myeloid differentiation were affected by normalization methodology such that using RMA, gcRMA, MAS5, and MBEI different numbers of differentially expressed genes were identified and only 12 of them were in common (Berkofsky-Fessler *et al.* 2004). It has been suggested that narrowing down the gene lists by excluding contradicting results from different methods might help increase reproducibility (Millenaar *et al.* 2006). Other studies compared the consistency among normalization methods at both the gene level and at the functional level (Raghavan *et al.* 2007). A better concordance was found at the functional biological level than at the gene level.

At the functional level, GO and KEGG Pathway annotations were consistent among the differentially expressed gene lists from different preprocessing methods although each list diffrered in terms of identity of the gene sets. For example, oxidative phosphorylation pathway was at the top of the lists for all there gene lists, which was also expected since hypoxia is an oxygen dependent condition. Although oxidative phosporylation pathway was the largest pathway in terms of significantly expressed genes, genes that were mapped to the pathway were not the same from one preprocessing method's list to another. In addition, top five GO categories that were shared by the largest number of genes were also consistent among the differentially expressed gene lists of each preprocessing method, for the level 3 categories of biological process, molecular function, and cellular component classes.

Finally, several studies suggest use of multiple normalization procedures at a time and importance of presence/absence thresholds to arrive at a more consistent gene list (Labbe *et al.* 2007). Williams *et al.* (2006) suggests guidelines to clean the data by

using platform specific normalizations, removal of absent data points, and inclusion of probesets that are robust to normalization procedures. Hubner *et al.* (2005) suggests that ranks in different gene lists may not be concordant but these lists may overlap to a greater extent when ranks not considered; Hubner *et al.* also goes against using probesets robust to normalization differences suggesting that RMA gives a lower FDR in their hands.

Our findings are in accord with obtaining a differentially expressed gene list that is significant in different preprocessing methods. Furthermore, a shorther list with invariant network parameters might help to select candidate genes for further study.

Gene Correlation Networks

Although differentially expressed genes are most of the time the major focus of microarray experiments, networks generated using correlated gene pairs are also beneficial for understanding gene function and gene interactions in a cellular context (Draghici 2003; Margolin *et al.* 2006). Tools of graph theory have been utilized for generating and analyzing gene expression correlation networks (Wang *et al.* 2006; Carter 2005).

For correlation networks, nodes represent genes and edges represent correlations. In order to compare different networks, topology measures are calculated to understand the overall characteristics of the network such as the clustering tendency of the network and the distribution of degrees of each node. Another crucial approach is to investigate genes in the network to see the relationships between groups of genes (Draghici 2003). With this approach, a gene's importance in a cell or its interactions with other genes can be identified using measures such as clustering coefficient, betweenness centrality, and degree distributions (Newman 2003; Verkhedkar *et al.* 2007). For example, Verkhedkar *et al.* (2007) showed that biological networks of *M. tuberculosis*, *M. leprae*, and *E. coli* exhibit a scale-free network structure.

However, it is difficult to assess a gene's or network's topological properties without a reference; thus networks are compared to each other or to random networks having similar or the same number of nodes and edges (Baralla *et al.* 2009; Newman 2003; de Haan *et al.* 2009; Wang *et al.* 2006; Verkhedkar *et al.* 2007). According to the

analysis of each network in this study, the network topology measures of interest such as clustering coefficient, average path length, and degree distribution values were significantly different from those generated from random counterparts.

In the present study, another reason that led us to use a randomly generated network for comparison with the hypoxia network was the lack of a public reference data for Affymetrix Zebrafish Genechip arrays, to the best of our knowledge. In addition, it is debatable that whether or not to use spike-in data since each cell type or condition impose its statistical properties on the data (Gyorffy *et al.* 2009; Shedden *et al.* 2005). Thus, spike-in data might be misleading for assessing the performance and efficiency of different preprocessing methods. In addition, since network topology parameters can differ from one network to another and topology measures cannot be interpreted without a reference network (Verkhedkar *et al.* 2007; Newman 2003), we used topology measures of random networks with similar nodes and degrees as a reference.

Interestingly, networks of union and intersection data were also significantly different from each other showing the amount of variability among genes that were significant in at least one of the preprocessed datasets. This finding suggested a potential benefit in using intersection data for a more conservative analysis and comparison of different preprocessing methods' networks. Robustness of using a gene list that is intersection of differentially expressed gene lists derived from differently preprocessed versions of a data is mentioned in a previous study (Rotter *et al.* 2008). They suggest that using such an intersection list would increase the likelihood of getting *de facto* differentially expressed genes in an experiment.

Comparison of Correlation Distributions

After the investigation of the basic characteristics of differently preprocessed data, correlation networks were constructed using only the positive Pearson correlation values. Two different datasets have been utilized for this purpose; union and intersection gene lists. The union gene list was constructed using the union of the significant gene lists of three preprocessing methods. The intersection gene list was also created using the intersection of these lists. Genes in intersection gene list expectedly were more correlated with each other than the ones in union gene list (Figures 11, 12). The reason behind this phenomenon might be the existence of

regulated transcription factors controlling the highly significant genes. In addition, a study from Rotter *et al.* (2008) also showed that significant genes that are verified from different approaches also have a higher likelihood of being *de facto* differentially expressed genes. Also, the link between the differentially expression of a gene and its connectivity in a protein-protein interaction network has been mentioned in the literature (Camargo *et al.* 2007). Thus, a list with genes more likely to be differentially expressed includes more connected pairs.

When comparing correlation distributions of each preprocessing method for union and intersection groups (Figures 13-16), RMA had the highest mean value of correlation coefficients whereas MAS5 had comparatively low values. This difference in terms of correlation coefficients has also been shown in the literature in different studies (Ploner *et al.* 2005; Lim *et al.* 2007).

One interesting pattern was the interchange of the mean of correlation values in union data around 0.45 which was not observed in intersection data. However, since the threshold for generating our network was 0.6, networks could be considered to be insensitive to the interchange between gcRMA and MAS5. Regarding the ranking of the mean values of correlation coefficients, it is mentioned in the literature that RMA and gcRMA might introduce false correlations between gene pairs (Lim *et al.* 2007). In the study of Lim *et al.* (2007), they have also shown that correlated gene pairs deducted from MAS5 preprocessed data was more related with actual protein-protein interaction data compared to RMA and gcRMA. Also, the effect of preprocessing methods in terms of correlation coefficient was in accord with our findings so that MAS5 had the lowest correlation values in randomly generated data. A possible explanation of the difference between the correlation distributions generated from different preprocessing methods has been given by Ploner *et al.* (2005). Accordingly different preprocessing methods efficiently remove the noise from different segments of data that differ in variability, i.e., low, medium, and high.

Overview of Network Topology Measures

Betweenness centrality: In this study, positive correlations were taken into account (Figures 13 and 14) and the slopes of the correlation curves were similar to each other

(Figures 15 and 16). Although the correlation structures were similar, network characteristics could not be deducted from those figures. So, betweenness centrality was measured for each network of union and intersection data. Random networks having the same number of genes and correlations with their counterparts were also generated for each preprocessing method's network in order to allow a better comparison for a reference. Betweenness centrality shows the number of times a gene is located on the shortest path of two different genes (Newman 2003). It can be thought as the popularity of a road that connects many cities (Lämmer *et al.* 2006). Genes with higher betweenness centrality values may have essential roles in regulation of other genes or establishing complexes (Hernández *et al.* 2007; Ahn *et al.* 2009). In this study we found that intersection data produced significant differences between preprocessing methods while no such difference could be observed for the union data.

Expectedly, each distribution highly significantly differed from its random network counterpart. Betweenness centrality is the measure of a gene's role of communication between two genes in a cellular process even the influence of that two genes is direct or indirect (Joy *et al.* 2005; Wang *et al.* 2009). Thus, it is expected that higher betweenness centrality measures would not be seen in random networks. Our findings are in parallel to the literature that betweenness centrality measures show a significant difference from randomly generated data (Hintze *et al.* 2008).

The differences between the intersection and union dataset in terms of network topology could be caused by the increased correlation coefficients due the the increased level of differential expression of gene lists. Thus, as networks get smaller and more clustered, due to the effects of preprocessing methods on network structure, distributions of betweenness centrality values cannot be generalized among the networks of differently preprocessing data. As shown in the study of Ploner *et al.* (2007), groups of genes with different variability have to be efficiently normalized by different preprocessing methods.

There also is an ongoing debate on the essentiality of a gene that whether essentiality can be represented better by betweenness centrality or node degree of that gene (Kar *et al.* 2009) (Yu *et al.* 2007; Zotenko *et al.* 2008). It might be interesting to compare

the essential and non-essential genes in their robustness to preprocessing algorithms using betweenness centrality measures in future studies.

Clustering coefficient: Clustering coefficient is a network topology measure to assess the tendency of establishing a cluster around a gene (Caretta-Cartozo *et al.* 2007). If a gene's clustering coefficient value is higher, its interaction with other gene groups and forming complexes in terms of interaction might be stronger (Newman 2003). In addition, clustering coefficient would reveal the importance of a gene in terms of a modular role in a cellular process (Ma'ayan 2009). Also, clustering coefficient would show the completeness and connectivity of a network (Horvath *et al.* 2008; Wang *et al.* 2009). From a different point of view, clustering coefficient would show the fault tolerance of a network; thus, it is expected to be higher than a random network as our study also shows (Supekar *et al.* 2008).

Looking at the results of the comparisons of distributions, it can be said that clustering coefficient is sensitive to the type of the preprocessing method (Tables 9 and 10). This is also in parallel to the study of Ploner *et al.* (2005) where the effects of different preprocessing methods on the correlation structure of genes have been assessed explained. Since every distribution is significantly different from each other, selecting an appropriate preprocessing method has a real importance in terms of getting a more reliable result. However, it seems to be almost impossible to find the appropriate preprocessing method without any reference; supportive information from RT-PCR yet would be invaluable for such a decision (Bolstad *et al.* 2003; Lim *et al.* 2007; Autio *et al.* 2009).

Degree distribution: Lastly, degree distribution has been the focus of this study in terms of network topology. Degree of a gene shows the number of correlations with other genes (Caretta-Cartozo *et al.* 2007). Interestingly, degree distributions of each preprocessing data were not significantly different from their random counterpart. However, it is mentioned in the literature that using only degree distribution for comparing different network topologies is not sufficient (Hormozdiari *et al.* 2007). This might be because random networks are generated based on the actual network's number of genes and correlations. However, apart from random networks, each degree distribution of actual data is significantly different from each other. This again

shows the differences of correlations among gene pairs changing from one preprocessing method to another one, as suggested by Ploner *et al* (2005).

As mentioned in the the betweenness centrality part, there is a debate on the importance of degree of a gene for estimating essentiality. However, it is not a resolved issue (Kar *et al.* 2009; Zotenko *et al.* 2008). Future studies might analyze the effect of preprocessing on essential and non-essential genes since their network structures are known to differ from each other.

Comparison of Network Topology Measures

In order to reveal the distribution of each network topology measure in response to the significance and fold change, various graphs have been plotted for intersection data (Figures 24-32). The reason for focusing on intersection data is that it leads to a more conservative comparative study when the main interest is on the significant genes between two groups of samples.

The difference of the upregulated and downregulated genes in terms of network topology measures has been studied in the literature (Wachi *et al.* 2005; Swindell 2008; Hernández *et al.* 2007). Accordingly, upregulated genes show weaker centrality values and are weakly connected (Swindell 2008). However, in another study, upregulated genes show higher correlation coefficients and better connectivity (Wachi *et al.* 2005). These differences between upregulated and downregulated genes might be either biological or due to type of the preprocessing method (Ploner *et al.* 2005). Although such differences have been observed, there is not a clear explanation for this phenomenon. We have also seen a difference between upregulated and downregulated gene groups in terms of network topology measures in our study focusing on positively correlated probesets.

In literature, different distributions of downregulated genes compared to upregulated ones in terms of network topology have been presented using gene expression and protein interaction networks (Hernández *et al.* 2007; Swindell 2008; Ahn *et al.* 2009). However, there is no clear explanation for this phenomenon (Hernández *et al.* 2007; Swindell 2008; Wachi *et al.* 2005). In our study, when network topology measures have been plotted against fold change, it was observed that in RMA and gcRMA

networks, downregulated genes have distinct values compared to upregulated genes (Figures 24, 26, 28). However, such a difference was not so obvious in MAS5 networks. This shows that network structure is affected by the preprocessing method. To the best of our knowledge, there is no comparative study in the literature on the effects of preprocessing methods on network topology to the best of our knowledge. Thus, our findings might reveal the disadvantage of MAS5 approach for reducing such effects from biological data. One possible explanation is mathematical; expression of a gene can be limitless whereas the minimum value of expression is zero. Thus, downregulated gene pairs might have higher correlation coefficients and interactions. Another explanation might be a biological one that downregulation might be a more controlled event than upregulation when a group of genes is shut down by the cell. Yet another explanation is that the phenomenon observed for the zebrafish hypoxia dataset might be condition and/or taxon-specific. Future studies should be performed to test the generality of our findings in other datasets from zebrafish and other organisms.

In order to compare interaction networks with random counterparts that are generated from actual data, 1932 genes (the number of genes in interaction data gen list) were randomly selected from each preprocessed data and corresponding networks were generated. Network topology measures of random networks seemed to be scattered randomly whereas the measures of significant genes in each random network showed similar distributions to the network measures of interaction data. In addition, network topology measures were symmetrical in terms of upregulated and downregulated genes when plotted by fold change. This might reveal the fact that the asymmetrical nature of network topology measures of significant genes for upregulated and downregulated lists occurred due the existence of a biological process (Figures 33 - 38). In terms of network topology measures, RMA, gcRMA and MAS5 do not significantly resemble each other. Low correlation coefficients of pair wise comparisons imply that each preprocessing method have distinctive network structure (Table 14). However, when networks of intersection data were compared using Spearman correlations RMA and gcRMA resembled each other with the highest correlation coefficient values (Table 13). The most distinctive pair is RMA-MAS5 where they have the lowest correlation coefficient value for each network topology measure. Even the lowest value is 0.55 which might show a true correlation in such a

network having ~2000 genes. This situation also suggests the usage of the intersection of significantly expressed gene lists from each preprocessing method since the network structure could be more conserved for those genes.

Genes having similar network topology measures

Although the general picture indicates that preprocessing methods radically alter the topology of correlation networks, genes that are the least affected by different preprocessing methods have been further investigated in this study. To find those genes, the same topologies for each preprocessing method's network were summed. For each network topology measures, sums of absolute differences were sorted and the least 20% was filtered. Accordingly, there were 91 and 81 genes for intersection and union data, respectively. Interestingly, although the number of genes in the intersection data had almost 40% of the genes of union data, the number of least-affected genes was higher. In addition, intersection of these gene lists had 31 genes common, suggesting that union and intersection data exhibited different characteristic properties in terms of network structure. It can be stressed again that each preprocessing method adds certain statistical properties to the data after the transformation (Draghici 2003; Irizarry, Hobbs, *et al.* 2003).

Investigating the GO categories of invariant genes in terms of network topology measures, it is intersting to observe biological categories that are relevant to a change in cellular profile as a result of hypoxia. Categories implying cell signaling and oxygen-related pathways show that the genes in those categories show a very robust expression and are insensitive to the preprocessing method. Further investigation of the expression profile and cellular characteristics of those genes might reveal the importance of using a network-based analysis to extract preprocessing-tolerant genes and pathways.

CHAPTER 6: FUTURE PERSPECTIVES

This study explored the effects of preprocessing algorithms on data distribution, differentially expressed genes, and network parameters using a Zebrafish GeneChip dataset. Accordingly, only positively correlated pairs of genes were considered to generate a network, for simplicity of the analysis. A detailed study considering negatively correlated gene pairs and all correlations would be meaningful in order to assess the effects of preprocessing on the network structure. Indeed, in a preliminary analysis of negatively correlated probeset pairs ($r<-0.6$) the pattern observed for positively correlated pairs was reversed (data not shown). Accordingly, upregulated genes instead of downregulated probesets exhibited greater connectivity and clustering coefficients. This suggests that one might lose the divergence between up- and downregulated probeset network topology if absolute correlation between probeset pairs is considered. Our study suggests that analyzing the gene networks for positively and negatively correlated pairs separately presents advantages for comparative studies. Furthermore, changing the threshold limit for the correlation coefficient as well as set p-value could affect the results and should be further studied. Effects of array quality might also be further tested on the observed patterns of divergence between up- and down-regulated genes.

In addition to the Pearson correlation method for identifying edges between genes as used in this thesis, usage of different distance metrics such as Spearman correlation or Euclidean distance might generate different networks. Thus, the extent of the effects of type of distance metrics used also can be evaluated in the future to understand the structural effects of such a choice on gene expression networks. This might be also beneficial to suggest a method for network generation that is least affecting the network structure.

In this study, a microarray dataset with 5 replicates in each group was used. Many microarray studies contain 3 or less replicates per group. The possibility of extending our method to experiment sets with low sample size could further be tested to define the limitations of the present method.

Detailed examination of the preprocessing methods could reveal the mechanistic effects of each preprocessing method. For example, changing the normalization or the

summarization method would also show the effect of the individual steps of preprocessing method on the structure of gene expression networks as well as the differentially expressed genes. Lastly, focusing on the sub-groups of data with different variabilities or coefficient of variations would reveal any regional effect that is caused by the preprocessing method.

In this study, we have applied a methodology minimizing the absolute differences between network parameters to obtain a robust set of probesets affected by hypoxia regardless of preprocessing method. Future studies might include extending this to usage of ranked parameter estimates.

In addition, applying the presented methodolgy to other zebrafish datasets from different tissues and/or pathologies would allow for its generalization. Applications to the datasets from other species would further contribute to the generalization of the findings in this thesis.

REFERENCES

Affymetrix. PLIER Technical Note.
http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf.

Array Design for the GeneChip Human Genome U133 Set.
http://www.abdn.ac.uk/ims/facilities/microarray/documents/hgu133_design_technote.pdf.

GeneChip® Zebrafish Genome Array.
http://www.affymetrix.com/support/technical/datasheets/zebrafish_datasheet.pdf.

Ahn, S., R. T. Wang, C. C. Park *et al*. 2009. Directed Mammalian Gene Regulatory Networks Using Expression and Comparative Genomic Hybridization Microarray Data from Radiation Hybrids. *PLoS Comput Biol* 5, no. 6 (June 12): e1000407. doi:10.1371/journal.pcbi.1000407.

Amatruda, J. F., and E. E. Patton. 2008. Genetic models of cancer in zebrafish. *International Review of Cell and Molecular Biology* 271: 1-34. doi:10.1016/S1937-6448(08)01201-X.

Anon. Gene Expression Omnibus (GEO) Main page. http://www.ncbi.nlm.nih.gov/geo/.

Anon. Autodesk - AutoCAD.
http://usa.autodesk.com/adsk/servlet/pc/index?siteID=123112&id=13779270.

Anon. BABELOMICS v3.1. http://babelomics.bioinfo.cipf.es/EntryPoint?loadForm=fatigo.

Anon. 2008. MatlabBGL 4.0. October.
http://www.stanford.edu/~dgleich/programs/matlab_bgl/.

Autio, R., S. Kilpinen, *et al.* 2009. Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. *BMC Bioinformatics* 10 Suppl 1: S24. doi:10.1186/1471-2105-10-S1-S24.

Bacha, J., J. S. Brodie *et al*. 2009. myGRN: a database and visualisation system for the storage and analysis of developmental genetic regulatory networks. *BMC Developmental Biology* 9: 33. doi:10.1186/1471-213X-9-33.

Baralla, A., W. I. Mentzen *et al*. 2009. Inferring gene networks: dream or nightmare? *Annals of the New York Academy of Sciences* 1158 (March): 246-256. doi:10.1111/j.1749-6632.2008.04099.x.

Barrat, A., and M. Weigt. 1999. On the properties of small-world network models. *cond-mat/9903411* (March 29). http://arxiv.org/abs/cond-mat/9903411.

Barrett, T., T. O. Suzek *et al.* 2005. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Research* 33, no. Database issue (January 1): D562-566. doi:10.1093/nar/gki022.

Berghmans, S., C. Jette *et al.* 2005. Making waves in cancer research: new models in the zebrafish. *BioTechniques* 39, no. 2 (August): 227-237.

Berkofsky-Fessler, W. D., M. J. McConnell *et al.* 2004. Identification of Genes Affected during Myeloid Differentiation Induced by Valproic Acid and Retinoic Acid: The Choice of Array Pre-Processing and Normalization Algorithm Has a Critical Effect on Results. *ASH Annual Meeting Abstracts* 104, no. 11 (November 16): 4207.

Beyene, J., P. Hu *et al.* 2007. Impact of normalization and filtering on linkage analysis of gene expression data. *BMC Proceedings* 1 Suppl 1: S150.

Bilban, M., L. K. Buehler *et al.* 2002. Normalizing DNA microarray data. *Current Issues in Molecular Biology* 4, no. 2 (April): 57-64.

Binder, H., S. Preibisch *et al.* 2010. Calibration of microarray gene-expression data. *Methods in Molecular Biology (Clifton, N.J.)* 576: 375-407. doi:10.1007/978-1-59745-545-9_20.

BioconductoR. Bioconductor. http://www.bioconductor.org/.

Biomart. Martview. http://www.biomart.org/.

Boes, T., and M. Neuhäuser. 2005. Normalization for Affymetrix GeneChips. *Methods of Information in Medicine* 44, no. 3: 414-7. doi:05030414.

Bolstad, B. M., R. A. Irizarry *et al.* 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)* 19, no. 2 (January 22): 185-93. doi:12538238.

Camargo, A., and F. Azuaje. 2007. Linking gene expression and functional network data in human heart failure. *PloS One* 2, no. 12: e1347. doi:10.1371/journal.pone.0001347.

Canales, R. D, Y. Luo *et al.* 2006. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology* 24, no. 9 (September): 1115-1122. doi:10.1038/nbt1236.

Caretta-Cartozo, C., P. D. L. Rios *et al.* 2007. Bottleneck genes and community structure in the cell cycle network of S. pombe. *PLoS Computational Biology* 3, no. 6 (June): e103. doi:10.1371/journal.pcbi.0030103.

Carter, G. W. 2005. Inferring network interactions within a cell. *Briefings in Bioinformatics* 6, no. 4 (December): 380-389.

Chiogna, M., M. S. Massa *et al.* 2009. A comparison on effects of normalisations in the detection of differentially expressed genes. *BMC Bioinformatics* 10, no. 1: 61. doi:10.1186/1471-2105-10-61.

Dooley, K., and L. I. Zon. 2000. Zebrafish: a model system for the study of human disease. *Current Opinion in Genetics & Development* 10, no. 3 (June): 252-256.

Draghici, S. 2003. *Data analysis tools for DNA microarrays*. Boca Raton: Chapman & Hall/CRC.

Freeman L. C. 1977. A Set of Measures of Centrality Based on Betweenness. *Sociometry* 40, no. 1 (March): 35-41.

Freeman, T. C., L. Goldovsky *et al.* 2007. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Computational Biology* 3, no. 10 (October): 2032-2042. doi:10.1371/journal.pcbi.0030206.

Fujita, A., J. R. Sato *et al.* 2006. Evaluating different methods of microarray data normalization. *BMC Bioinformatics* 7: 469. doi:10.1186/1471-2105-7-469.

Geller, S. C, J. P. Gregg *et al.* 2003. Transformation and normalization of oligonucleotide microarray data. *Bioinformatics (Oxford, England)* 19, no. 14 (September 22): 1817-23. doi:14512353.

Giles, P. J., and D. Kipling. 2003. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics (Oxford, England)* 19, no. 17 (November 22): 2254-62. doi:14630654.

Gonzalez-Nunez, V., and R. E. Rodríguez. 2009. The zebrafish: a model to study the endogenous mechanisms of pain. *ILAR Journal / National Research Council, Institute of Laboratory Animal Resources* 50, no. 4: 373-386.

Grewal, A., P. Lambert *et al.* 2007. Analysis of expression data: an overview. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al* Chapter 7 (March): Unit 7.1. doi:10.1002/0471250953.bi0701s17.

Gumbel, E. J. 1958. *Statistics of Extremes*. Columbia Univ Pr, June.

Gyorffy, B., B. Molnar *et al.* 2009. Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PloS One* 4, no. 5: e5645. doi:10.1371/journal.pone.0005645.

de Haan, W., Y. A. L. Pijnenburg *et al.* 2009. Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory. *BMC Neuroscience* 10: 101. doi:10.1186/1471-2202-10-101.

Harr, B., and C. Schlötterer. 2006. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Research* 34, no. 2: e8. doi:10.1093/nar/gnj010.

Hernández, P., J. Huerta-Cepas *et al.* 2007. Evidence for systems-level molecular mechanisms of tumorigenesis. *BMC Genomics* 8: 185. doi:10.1186/1471-2164-8-185.

Hill, A. A., E. L. Brown *et al.* 2001. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biology* 2, no. 12: RESEARCH0055.

Hintze, A., and C. Adami. 2008. Evolution of complex modular biological networks. *PLoS Computational Biology* 4, no. 2 (February): e23. doi:10.1371/journal.pcbi.0040023.

Hormozdiari, F., P. Berenbrink *et al.* 2007. Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution. *PLoS Computational Biology* 3, no. 7 (July): e118. doi:10.1371/journal.pcbi.0030118.

Horvath, S., and J. Dong. 2008. Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology* 4, no. 8: e1000117. doi:10.1371/journal.pcbi.1000117.

Hua, Y. J., K. Tu *et al.* 2008. Comparison of normalization methods with microRNA microarray. *Genomics* 92, no. 2 (August): 122-128. doi:10.1016/j.ygeno.2008.04.002.

Hubbell, E., W. M. Liu *et al.* 2002. Robust estimators for expression analysis. *Bioinformatics (Oxford, England)* 18, no. 12 (December): 1585-1592.

Hubner, N., C. A. Wallace *et al.* 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* 37, no. 3 (March): 243-253. doi:10.1038/ng1522.

Irizarry, R. A., B. M. Bolstad *et al.* 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* 31, no. 4 (February 15): e15. doi:12582260.

Irizarry, R. A., B. Hobbs *et al.* 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)* 4, no. 2 (April): 249-64. doi:12925520.

Joy, M. P., A. Brock *et al.* 2005. High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine & Biotechnology* 2005, no. 2 (June 30): 96-103. doi:10.1155/JBB.2005.96.

Jupiter, D., H. Chen *et al.* 2009. STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data. *BMC Bioinformatics* 10: 332. doi:10.1186/1471-2105-10-332.

Kadota, K., Y. Nakai *et al.* 2008. A weighted average difference method for detecting differentially expressed genes from microarray data. *Algorithms for Molecular Biology: AMB* 3: 8. doi:10.1186/1748-7188-3-8.

Kadota, K., Y. Nakai *et al.* 2009. Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity. *Algorithms for Molecular Biology: AMB* 4: 7. doi:10.1186/1748-7188-4-7.

Kar, G., A. Gursoy *et al.* 2009. Human cancer protein-protein interaction network: a structural perspective. *PLoS Computational Biology* 5, no. 12 (December): e1000601. doi:10.1371/journal.pcbi.1000601.

Katz, S., R. A. Irizarry *et al.* 2006. A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC Bioinformatics* 7: 464. doi:10.1186/1471-2105-7-464.

Kreil, D. P., and R. R. Russell. 2005. There is no silver bullet--a guide to low-level data transforms and normalisation methods for microarray data. *Briefings in Bioinformatics* 6, no. 1 (March): 86-97.

Labbe, A., M. P. Roth *et al.* 2007. Impact of gene expression data pre-processing on expression quantitative trait locus mapping. *BMC Proceedings* 1, no. Suppl 1: S153.

Lämmer, S., B. Gehlsen *et al.* 2006. Scaling laws in the spatial structure of urban road networks. *Physica A: Statistical Mechanics and its Applications* 363, no. 1 (April 15): 89-95. doi:10.1016/j.physa.2006.01.051.

Li, C., and W. H. Wong. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* 2, no. 8: RESEARCH0032.

Lim, W. K., K. Wang *et al.* 2007. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics (Oxford, England)* 23, no. 13 (July 1): i282-8. doi:23/13/i282.

Liu, W. M., R. Li *et al.* 2006. PQN and DQN: algorithms for expression microarrays. *Journal of Theoretical Biology* 243, no. 2 (November 21): 273-278. doi:10.1016/j.jtbi.2006.06.017.

Ma'ayan, A. 2009. Insights into the organization of biochemical regulatory networks using graph theory analyses. *The Journal of Biological Chemistry* 284, no. 9 (February 27): 5451-5455. doi:10.1074/jbc.R800056200.

Margolin, A. A., I. Nemenman *et al.* 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* 7 Suppl 1: S7.

Marques, I. J., J. T. D. Leito *et al.* 2008. Transcriptome analysis of the response to chronic constant hypoxia in zebrafish hearts. *Journal of Comparative Physiology. B, Biochemical, Systemic, and Environmental Physiology* 178, no. 1 (January): 77-92. doi:10.1007/s00360-007-0201-4.

Millenaar, F. F., J. Okyere *et al.* 2006. How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics* 7: 137. doi:1471-2105-7-137.

Murie, C., O. Woody *et al.* 2009. Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics* 10: 45. doi:10.1186/1471-2105-10-45.

Newman, M. E. J. 2003. The structure and function of complex networks. *cond-mat/0303516* (March 25). http://arxiv.org/abs/cond-mat/0303516.

Pease, A. C., D. Solas *et al.* 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America* 91, no. 11 (May 24): 5022-5026.

Pepper, S. D., E. K. Saunders *et al.* 2007. The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics* 8: 273. doi:10.1186/1471-2105-8-273.

Ploner, A., L. D. Miller *et al.* 2005. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC Bioinformatics* 6: 80. doi:10.1186/1471-2105-6-80.

Qin, L. X., R. Beyer *et al.* 2006. Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics* 7, no. 1: 23. doi:10.1186/1471-2105-7-23.

Qiu, X, A. I. Brooks *et al.* 2005. The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* 6: 120. doi:10.1186/1471-2105-6-120.

Quackenbush, J. 2002. Microarray data normalization and transformation. *Nature Genetics* 32 Suppl (December): 496-501. doi:10.1038/ng1032.

Raghavan, N., A. M. I. M. De Bondt *et al.* 2007. The high-level similarity of some disparate gene expression measures. *Bioinformatics* 23, no. 22 (November 15): 3032-3038. doi:10.1093/bioinformatics/btm448.

Reverter, A., W. Barris *et al.* 2005. Validation of alternative methods of data normalization in gene co-expression studies. *Bioinformatics (Oxford, England)* 21, no. 7 (April 1): 1112-20. doi:bti124.

Rotter, A., M. Hren *et al.* 2008. Finding differentially expressed genes in two-channel DNA microarray datasets: how to increase reliability of data preprocessing. *Omics: A Journal of Integrative Biology* 12, no. 3 (September): 171-182. doi:10.1089/omi.2008.0032.

Schena, M., D. Shalon *et al.* 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)* 270, no. 5235 (October 20): 467-470.

Schuster, E. F., E. Blanc *et al.* 2007. Estimation and correction of non-specific binding in a large-scale spike-in experiment. *Genome Biology* 8, no. 6: R126. doi:10.1186/gb-2007-8-6-r126.

Seidel, M., and R. Niessner. 2008. Automated analytical microarrays: a critical review. *Analytical and Bioanalytical Chemistry* 391, no. 5 (July): 1521-44. doi:10.1007/s00216-008-2039-3.

Selga, E., C. Oleaga *et al.* 2009. Networking of differentially expressed genes in human cancer cells resistant to methotrexate. *Genome Medicine* 1, no. 9: 83. doi:10.1186/gm83.

Semenza, G. L. 2001. Hypoxia-inducible factor 1: oxygen homeostasis and disease pathophysiology. *Trends in Molecular Medicine* 7, no. 8 (August): 345-350.

Shedden, K., W. Chen *et al.* 2005. Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics* 6: 26. doi:1471-2105-6-26.

Smyth, G. K., and T. Speed. 2003. Normalization of cDNA microarray data. *Methods (San Diego, Calif.)* 31, no. 4 (December): 265-273.

Sorathiya, A., T. Jucikas *et al.* 2009. Searching for Glycomics Role in Stem Cell Development. In *Computational Intelligence Methods for Bioinformatics and Biostatistics: 5th International Meeting, CIBB 2008 Vietri sul Mare, Italy, October 3-4, 2008 Revised Selected Papers*, 198-209. Springer-Verlag. http://portal.acm.org/citation.cfm?id=1574987.

Sporns, O. , C. J. Honey *et al.* 2007. Identification and Classification of Hubs in Brain Networks. *PLoS ONE* 2, no. 10 (October 17): e1049. doi:10.1371/journal.pone.0001049.

Stecyk, J. A. W., K. O. Stensløkken *et al.* 2004. Maintained cardiac pumping in anoxic crucian carp. *Science (New York, N.Y.)* 306, no. 5693 (October 1): 77. doi:10.1126/science.1100763.

Steinhoff, C., and M. Vingron. 2006. Normalization and quantification of differential expression in gene expression microarrays. *Briefings in Bioinformatics* 7, no. 2 (June): 166-177. doi:10.1093/bib/bbl002.

Strogatz, S. H. 2001. Exploring complex networks. *Nature* 410, no. 6825 (March 8): 268-276. doi:10.1038/35065725.

Supekar, K., V. Menon *et al.* 2008. Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. *PLoS Computational Biology* 4, no. 6 (June): e1000100. doi:10.1371/journal.pcbi.1000100.

Swindell, W. R. 2008. Genes regulated by caloric restriction have unique roles within transcriptional networks. *Mechanisms of Ageing and Development* 129, no. 10 (October): 580-592. doi:10.1016/j.mad.2008.06.001.

Thorgeirsson, S. S., J. S. Lee *et al.* 2006. Functional genomics of hepatocellular carcinoma. *Hepatology (Baltimore, Md.)* 43, no. 2 Suppl 1 (February): S145-150. doi:10.1002/hep.21063.

Vardhanabhuti, S., S. J. Blakemore *et al.* 2006. A comparison of statistical tests for detecting differential expression using Affymetrix oligonucleotide microarrays. *Omics: A Journal of Integrative Biology* 10, no. 4: 555-566. doi:10.1089/omi.2006.10.555.

Verhaak, R. G. W., F. J. T. Staal *et al.* 2006. The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies. *BMC Bioinformatics* 7: 105. doi:10.1186/1471-2105-7-105.

Verkhedkar, K. D., K. Raman *et al.* 2007. Metabolome based reaction graphs of M. tuberculosis and M. leprae: a comparative network analysis. *PloS One* 2, no. 9: e881. doi:10.1371/journal.pone.0000881.

Wachi, S., K. Yoneda *et al.* 2005. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics (Oxford, England)* 21, no. 23 (December 1): 4205-4208. doi:10.1093/bioinformatics/bti688.

Wang, J., S. Zhang *et al.* 2009. Disease-aging network reveals significant roles of aging genes in connecting genetic diseases. *PLoS Computational Biology* 5, no. 9 (September): e1000521. doi:10.1371/journal.pcbi.1000521.

Wang, Y., T. Joshi *et al.* 2006. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics (Oxford, England)* 22, no. 19 (October 1): 2413-2420. doi:10.1093/bioinformatics/btl396.

Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, no. 6684 (June 4): 440-442. doi:10.1038/30918.

Webb, K. J., W.H. Norton *et al.* 2009. Zebrafish reward mutants reveal novel transcripts mediating the behavioral effects of amphetamine. *Genome Biology* 10, no. 7: R81. doi:10.1186/gb-2009-10-7-r81.

Williams, R. B. H., C. J. Cotsapas *et al.* 2006. Normalization procedures and detection of linkage signal in genetical-genomics experiments. *Nature Genetics* 38, no. 8 (August): 855-856; author reply 856-859. doi:10.1038/ng0806-855.

Wiltgen, M., and G. P. Tilz. 2007. DNA microarray analysis: principles and clinical impact. *Hematology (Amsterdam, Netherlands)* 12, no. 4 (August): 271-287. doi:10.1080/10245330701283967.

Yauk, C. L., and M. L. Berndt. 2007. Review of the literature examining the correlation among DNA microarray technologies. *Environmental and Molecular Mutagenesis* 48, no. 5 (June): 380-394. doi:10.1002/em.20290.

Yu, H., P. M. Kim *et al.* 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Computational Biology* 3, no. 4 (April 20): e59. doi:10.1371/journal.pcbi.0030059.

Zakharkin, S. O., K. Kim *et al.* 2005. Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics* 6: 214. doi:10.1186/1471-2105-6-214.

Zhijin W., R. A. Irizarry *et al.* 2004. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* 99, no. 468 (December 1): 909-917.

Zotenko, E., J. Mestre *et al.* 2008. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Computational Biology* 4, no. 8: e1000140. doi:10.1371/journal.pcbi.1000140.

APPENDIX A

R script for preparing the preprocessed data is given below:

```
affydata = ReadAffy();  # read .CEL files from the current directory
outputfileprefix = 'GSE4989'; # prepare output file name for saving

mas5data = mas5(affydata)      # apply MAS5 algorithm to affymetrix
data
exprsvalues = exprs(mas5data) # get expression values of the
structured data
log2data = log(exprsvalues, 2)      # calculate log2 values of data
outputfile = paste(outputfileprefix, '.MAS5.txt', sep = "", collapse
= NULL)
write.table(log2data, file=outputfile, quote=F, sep="\t")   #save
data

rmadata = rma(affydata) # apply RMA algorithm to affymetrix data
log2data= exprs(rmadata)# get expression values of the structured
data
outputfile = paste(outputfileprefix, '.RMA.txt', sep = "", collapse =
NULL)
write.table(log2data, file=outputfile, quote=F, sep="\t")   #save
data

gcrmadata = gcrma(affydata)   # apply gcRMA algorithm to affymetrix
data
log2data= exprs(gcrmadata)    # get expression values of the
structured data
outputfile = paste(outputfileprefix, '.gcRMA.txt', sep = "", collapse
= NULL)
write.table(log2data, file=outputfile, quote=F, sep="\t")   #save
data
```

APPENDIX B

MatLab script applying t-test for each differently preprocessed data is given below:

```
function [pvalues ids
data]=aro_ttest(xls_file,controlgroup,experimentgroup)
      %read excel data and do a ttest to get pvalues
      [data text] = xlsread(xls_file);    % read excel data
      controldata = data(:, controlgroup);
      experimentdata = data(:, experimentgroup);
      ids = text(2:end, 1);   % assign probe names
      pvalues = mattest(controldata, experimentdata); % apply t-test



hypoxiaindex = [1 3 5 8 9];   % column index of conditions with
hypoxia
normoxiaindex = [2 4 6 7 10]; % column index of conditions with
normoxia

[rmapvalues ids rmadata] = aro_ttest('GSE4989.RMA.xls',
normoxiaindex,hypoxiaindex);
[gcrmapvalues ids gcrmadata] =
aro_ttest('GSE4989.gcRMA.xls',normoxiaindex,hypoxiaindex);
[mas5pvalues ids mas5data] =
aro_ttest('GSE4989.MAS5.xls',normoxiaindex,hypoxiaindex);


% get significant probe set lists from each differently preprocessed
dataset
rmaindex = rmapvalues <= 0.05;
rmaprobes = ids(rmaindex);

gcrmaindex = gcrmapvalues <= 0.05;
gcrmaprobes = ids(gcrmaindex);

mas5index = mas5pvalues <= 0.05;
mas5probes = ids(mas5index);


% get union of differentially expressed genes of each preprocessed
data pair
unionprobes = union(rmaprobes, union(gcrmaprobes, mas5probes));
unionrma_gcrma = union(rmaprobes, gcrmaprobes);
unionrma_mas5 = union(rmaprobes, mas5probes);
uniongcrma_mas5 = union(gcrmaprobes, mas5probes);

%get intersection of differentially expressed genes of each
preprocessed data pair
intersectprobes = intersect(rmaprobes,intersect(gcrmaprobes,
mas5probes));
intersectrma_gcrma = intersect(rmaprobes, gcrmaprobes);
intersectrma_mas5 = intersect(rmaprobes, mas5probes);
intersectgcrma_mas5 = intersect(gcrmaprobes, mas5probes);
```

APPENDIX C

MatLab script for generating networks of each union and intersection data for each differently preprocessed data is given below:

```
function corrnet = aro_corrnet(data, message, limit)
      %generates 0-1 correlation network
      disp(['Generating network: ' message]);
      datasize = size(data,1);
      tempcorr = corr(data');
      corrnet = tempcorr;
      corrnet(tempcorr>=limit) = 1;
      corrnet (tempcorr<limit) = 0;
      corrnet(isnan(corrnet)) = 0;


function randmat = aro_randmatrix(rt)
      %rt is the original matrix to be randomized
      merged = [];
      lenrt = length(rt)
      for i = 1:(lenrt-1)
          merged = [merged rt(i, (i+1):end)];
      end
      merged = merged(randperm(length(merged)));
      randmat = zeros(lenrt, lenrt);
      for i = 1:(lenrt-1)
          curr = merged(1:(lenrt-i));
          merged = merged((lenrt-i+1):end);
          randmat(i,(i+1):end) = curr;
      end
      randmat = randmat + randmat';
      for i = 1:lenrt
          randmat(i,i) = 1;
      end


rlimit=0.6; % setting threshold of r-value for network generation
probes = ids;

%generating correlation networks using union of gene lists
[nosense1 unionprobesindex nosense2] = intersect(probes,
unionprobes);
unioncontrolprobes = probes(unionprobesindex);

%filtering union data from each whole-data
unionrmadata = rmadata(unionprobesindex, :);
uniongcrmadata = gcrmadata(unionprobesindex, :);
unionmas5data = mas5data(unionprobesindex, :);

%generating correlation networks and random networks
unionrmanet = aro_corrnet(unionrmadata, 'Union RMA', rlimit);
unionrmanetsparse = sparse(double(unionrmanet));
unionrmanetsparserandom =
sparse(double(aro_randmatrix(unionrmanet)));
```

```
uniongcrmanet = aro_corrnet(uniongcrmadata, 'Union gcRMA', rlimit);
uniongcrmanetsparse = sparse(double(uniongcrmanet));
uniongcrmanetsparserandom =
sparse(double(aro_randmatrix(uniongcrmanet)));

unionmas5net = aro_corrnet(unionmas5data, 'Union MAS5', rlimit);
unionmas5netsparse = sparse(double(unionmas5net));
unionmas5netsparserandom =
sparse(double(aro_randmatrix(unionmas5net)));

%generating correlation networks using intersection of gene lists
[nosense1 intersectprobesindex nosense2] = intersect(probes,
intersectprobes);
intersectcontrolprobes = probes(intersectprobesindex);
% filtering intersection data from each whole-data
intersectrmadata = rmadata(intersectprobesindex, :);
intersectgcrmadata = gcrmadata(intersectprobesindex, :);
intersectmas5data = mas5data(intersectprobesindex, :);

%generating correlation networks and random networks
intersectrmanet = aro_corrnet(intersectrmadata, 'Intersect RMA',
rlimit);
intersectrmanetsparse = sparse(double(intersectrmanet));
intersectrmanetsparserandom =
sparse(double(aro_randmatrix(intersectrmanet)));


intersectgcrmanet = aro_corrnet(intersectgcrmadata, 'Intersect
gcRMA', rlimit);
intersectgcrmanetsparse = sparse(double(intersectgcrmanet));
intersectgcrmanetsparserandom =
sparse(double(aro_randmatrix(intersectgcrmanet)));

intersectmas5net = aro_corrnet(intersectmas5data, 'Intersect MAS5',
rlimit);
intersectmas5netsparse = sparse(double(intersectmas5net));
intersectmas5netsparserandom =
sparse(double(aro_randmatrix(intersectmas5net)));
```

APPENDIX D

MatLab Script for calculating network topology measures is given below:

```
% calculate clustering coefficient for each intersection data
disp('CC for intersect RMA')
ccintrma = clustering_coefficients(intersectrmanetsparse);
ccintrmarand = clustering_coefficients(intersectrmanetsparserandom);

disp('CC for intersect gcRMA')
ccintgcrma = clustering_coefficients(intersectgcrmanetsparse);
ccintgcrmarand =
clustering_coefficients(intersectgcrmanetsparserandom);

disp('CC for intersect MAS5')
ccintmas5 = clustering_coefficients(intersectmas5netsparse);
ccintmas5rand =
clustering_coefficients(intersectmas5netsparserandom);


% calculate betweenness centrality for each intersection data
disp('BC for intersect RMA')
bcintrma = betweenness_centrality(intersectrmanetsparse);
bcintrmarand = betweenness_centrality(intersectrmanetsparserandom);

disp('BC for intersect gcRMA')
bcintgcrma = betweenness_centrality(intersectgcrmanetsparse);
bcintgcrmarand =
betweenness_centrality(intersectgcrmanetsparserandom);

disp('BC for intersect MAS5')
bcintmas5 = betweenness_centrality(intersectmas5netsparse);
bcintmas5rand = betweenness_centrality(intersectmas5netsparserandom);


% calculate degree distribution for each intersection data
disp('DD for intersect RMA')
ddintrma = sum(intersectrmanetsparse);
ddintrmarand = sum(intersectrmanetsparserandom);

disp('DD for intersect gcRMA')
ddintgcrma = sum(intersectgcrmanetsparse);
ddintgcrmarand = sum(intersectgcrmanetsparserandom);

disp('DD for intersect MAS5')
ddintmas5 = sum(intersectmas5netsparse);
ddintmas5rand = sum(intersectmas5netsparserandom);


% calculate clustering coefficient for each union data
disp('CC for union RMA')
ccunirma = clustering_coefficients(unionrmanetsparse);
ccunirmarand = clustering_coefficients(unionrmanetsparserandom);

disp('CC for union gcRMA')
ccunigcrma = clustering_coefficients(uniongcrmanetsparse);
```

```
ccunigcrmarand = clustering_coefficients(uniongcrmanetsparserandom);

disp('CC for union MAS5')
ccunimas5 = clustering_coefficients(unionmas5netsparse);
ccunimas5rand = clustering_coefficients(unionmas5netsparserandom);


% calculate betweenness centrality for each union data
disp('BC for union RMA')
bcunirma = betweenness_centrality(unionrmanetsparse);
bcunirmarand = betweenness_centrality(unionrmanetsparserandom);

disp('BC for union gcRMA')
bcunigcrma = betweenness_centrality(uniongcrmanetsparse);
bcunigcrmarand = betweenness_centrality(uniongcrmanetsparserandom);

disp('BC for union MAS5')
bcunimas5 = betweenness_centrality(unionmas5netsparse);
bcunimas5rand = betweenness_centrality(unionmas5netsparserandom);



% calculate degree distribution for each union data
disp('DD for union RMA')
ddunirma = sum(unionrmanetsparse);
ddunirmarand = sum(unionrmanetsparserandom);

disp('DD for union gcRMA')
ddunigcrma = sum(uniongcrmanetsparse);
ddunigcrmarand = sum(uniongcrmanetsparserandom);

disp('DD for union MAS5')
ddunimas5 = sum(unionmas5netsparse);
ddunimas5rand = sum(unionmas5netsparserandom);


% plotting figures for intersection data
figure();
boxplot([bcintrma bcintrmarand bcintgcrma bcintgcrmarand bcintmas5
bcintmas5rand])
set(gca, 'XTickLabel', [{'RMA'} {'RMA-Random'} {'gcRMA'} {'gcRMA-
Random'} {'MAS5'} {'MAS5-Random'} ])
xlabel('')
ylabel('Betweennes Centrality')
title('Intersection Data')

figure();
boxplot([ccintrma ccintrmarand ccintgcrma ccintgcrmarand ccintmas5
ccintmas5rand])
set(gca, 'XTickLabel', [{'RMA'} {'RMA-Random'} {'gcRMA'} {'gcRMA-
Random'} {'MAS5'} {'MAS5-Random'} ])
xlabel('')
ylabel('Clustering Coefficient')
title('Intersection Data')

figure();
boxplot([ddintrma' ddintrmarand' ddintgcrma' ddintgcrmarand'
ddintmas5' ddintmas5rand'])
set(gca, 'XTickLabel', [{'RMA'} {'RMA-Random'} {'gcRMA'} {'gcRMA-
Random'} {'MAS5'} {'MAS5-Random'} ])
xlabel('')
```

```
ylabel('Degree Distribution')
title('Intersection Data')


% plotting figures for union data
figure();
boxplot([bcunirma bcunirmarand bcunigcrma bcunigcrmarand bcunimas5
bcunimas5rand])
set(gca, 'XTickLabel', [{'RMA'} {'RMA-Random'} {'gcRMA'} {'gcRMA-
Random'} {'MAS5'} {'MAS5-Random'} ])
xlabel('')
ylabel('Betweennes Centrality')
title('Union Data')

figure();
boxplot([ccunirma ccunirmarand ccunigcrma ccunigcrmarand ccunimas5
ccunimas5rand])
set(gca, 'XTickLabel', [{'RMA'} {'RMA-Random'} {'gcRMA'} {'gcRMA-
Random'} {'MAS5'} {'MAS5-Random'} ])
xlabel('')
ylabel('Clustering Coefficient')
title('Union Data')

figure();
boxplot([ddunirma' ddunirmarand' ddunigcrma' ddunigcrmarand'
ddunimas5' ddunimas5rand'])
set(gca, 'XTickLabel', [{'RMA'} {'RMA-Random'} {'gcRMA'} {'gcRMA-
Random'} {'MAS5'} {'MAS5-Random'} ])
xlabel('')
ylabel('Degree Distribution')
title('Union Data')
```

APPENDIX E

MatLab Script for generating a random correlation matrix with the same number of nodes and edges of a given network is given below:

```
function randmat = aro_randmatrix(rt)
    %rt is the original matrix to be randomized
    merged = [];
    lenrt = length(rt);
    for i = 1:(lenrt-1)
        merged = [merged rt(i, (i+1):end)];
    end
    merged = merged(randperm(length(merged)));
    randmat = zeros(lenrt, lenrt);
    for i = 1:(lenrt-1)
        curr = merged(1:(lenrt-i));
        merged = merged((lenrt-i+1):end);
        randmat(i,(i+1):end) = curr;
    end
    randmat = randmat + randmat';
    for i = 1:lenrt
        randmat(i,i) = 1;
    end
```

APPENDIX F

Matlab scripts for listing the least-changed genes in each network topology measure, for both union and intersection data.

```
function sorted = aro_mergeandsort(probenames, rmavalues,
gcrmavalues, mas5values)
    rm = rmavalues - mas5values;
    rg = rmavalues - gcrmavalues;
    mg = mas5values - gcrmavalues;
    rg = abs(rg);
    rm = abs(rm);
    mg = abs(mg);
    sumofmeasures = rg + rm + mg;
    [empty sortindex] = sort(sumofmeasures);
    sorted = probenames(sortindex);


bcunisorted = aro_mergeandsort(unionprobes, bcunirma, bcunigcrma,
bcunimas5);
bcintsorted = aro_mergeandsort(intersectprobes, bcintrma, bcintgcrma,
bcintmas5);

ccunisorted = aro_mergeandsort(unionprobes, ccunirma, ccunigcrma,
ccunimas5);
ccintsorted = aro_mergeandsort(intersectprobes, ccintrma, ccintgcrma,
ccintmas5);

ddunisorted = aro_mergeandsort(unionprobes, ddunirma, ddunigcrma,
ddunimas5);
ddintsorted = aro_mergeandsort(intersectprobes, ddintrma, ddintgcrma,
ddintmas5);


n = round(length(unionprobes) * 0.20); %top 20%
unicommon = intersect(bcunisorted(1:n), intersect(ccunisorted(1:n),
ddunisorted(1:n)));
n = round(length(intersectprobes) * 0.20); % top 20%
intcommon = intersect(bcintsorted(1:n), intersect(ccintsorted(1:n),
ddintsorted(1:n)));
```

APPENDIX G

Following scripts are utilized for Spearman correlation calculation for each network topology parameters of pairs of different preprocessing datasets.

```
corr(ccrmarand, ccmas5rand, 'type', 'Spearman')
corr(ccrmarand, ccgcrmarand, 'type', 'Spearman')
corr(ccgcrmarand, ccmas5rand, 'type', 'Spearman')

corr(bcrmarand, bcmas5rand, 'type', 'Spearman')
corr(bcrmarand, bcgcrmarand, 'type', 'Spearman')
corr(bcgcrmarand, bcmas5rand, 'type', 'Spearman')

corr(ddrmarand', ddmas5rand', 'type', 'Spearman')
corr(ddrmarand', ddgcrmarand', 'type', 'Spearman')
corr(ddgcrmarand', ddmas5rand', 'type', 'Spearman')
```

APPENDIX H

Following functions are used to generate plots of each network topology measure versus p-value and fold change.

Plotting random data:

```
%calculating network topology measures for random data with the size
of
%intersection data
lenrandom = 1:length(rmadata);
randomindex = lenrandom(randperm(length(intersectprobes)));
randomprobes = probes(randomindex);

rmarand = rmadata(randomindex, :);
gcrmarand = gcrmadata(randomindex, :);
mas5rand = mas5data(randomindex, :);


rmarandpval = mattest(rmarand(:,normoxiaindex),
rmarand(:,hypoxiaindex));
rmarandfc = mean(rmarand(:,hypoxiaindex)') -
mean(rmarand(:,normoxiaindex)');

gcrmarandpval = mattest(gcrmarand(:,normoxiaindex),
gcrmarand(:,hypoxiaindex));
gcrmarandfc = mean(gcrmarand(:,hypoxiaindex)') -
mean(gcrmarand(:,normoxiaindex)');

mas5randpval = mattest(mas5rand(:,normoxiaindex),
mas5rand(:,hypoxiaindex));
mas5randfc = mean(mas5rand(:,hypoxiaindex)') -
mean(mas5rand(:,normoxiaindex)');



rmarandnet = aro_corrnet(rmarand, 'random-RMA', rlimit);
rmarandnetsparse = sparse(double(rmarandnet));
gcrmarandnet = aro_corrnet(gcrmarand, 'random-gcRMA', rlimit);
gcrmarandnetsparse = sparse(double(gcrmarandnet));
mas5randnet = aro_corrnet(mas5rand, 'random-MAS5', rlimit);
mas5randnetsparse = sparse(double(mas5randnet));


disp('CC for random RMA')
ccrmarand = clustering_coefficients(rmarandnetsparse);
disp('BC for random RMA')
bcrmarand = betweenness_centrality(rmarandnetsparse);
disp('DD for random RMA')
ddrmarand = sum(rmarandnetsparse);


disp('CC for random gcRMA')
ccgcrmarand = clustering_coefficients(gcrmarandnetsparse);
disp('BC for random gcRMA')
bcgcrmarand = betweenness_centrality(gcrmarandnetsparse);
```

```
disp('DD for random gcRMA')
ddgcrmarand = sum(gcrmarandnetsparse);


disp('CC for random MAS5')
ccmas5rand = clustering_coefficients(mas5randnetsparse);
disp('BC for random MAS5')
bcmas5rand = betweenness_centrality(mas5randnetsparse);
disp('DD for random MAS5')
ddmas5rand = sum(mas5randnetsparse);


%PLOTTING FIGURES

%fc vs. cc
figure;
subplot(1,3,1)
rmarandfcp = rmarandfc > 0;
rmarandfcn = rmarandfc < 0;
plot(rmarandfc(rmarandfcp), ccrmarand(rmarandfcp), '.r',
'MarkerSize', 5);
hold on;
plot(rmarandfc(rmarandfcn), ccrmarand(rmarandfcn), '.b',
'MarkerSize', 5);
title('RMA random - FC vs CC')
xlim([-10 10])
ylim([0 1])

%fc vs. cc
subplot(1,3,2)
gcrmarandfcp = gcrmarandfc > 0;
gcrmarandfcn = gcrmarandfc < 0;
plot(gcrmarandfc(gcrmarandfcp), ccgcrmarand(gcrmarandfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmarandfc(gcrmarandfcn), ccgcrmarand(gcrmarandfcn), '.b',
'MarkerSize', 5);
title('gcRMA random - FC vs CC')
xlim([-10 10])
ylim([0 1])

%fc vs. cc
subplot(1,3,3)
mas5randfcp = mas5randfc > 0;
mas5randfcn = mas5randfc < 0;
plot(mas5randfc(mas5randfcp), ccmas5rand(mas5randfcp), '.r',
'MarkerSize', 5);
hold on;
plot(mas5randfc(mas5randfcn), ccmas5rand(mas5randfcn), '.b',
'MarkerSize', 5);
title('MAS5 random - FC vs CC')
xlim([-10 10])
ylim([0 1])


%fc vs. bc
figure;
subplot(1,3,1)
hold on;
plot(rmarandfc(rmarandfcp), bcrmarand(rmarandfcp), '.r',
'MarkerSize', 5);
```

```
plot(rmarandfc(rmarandfcn), bcrmarand(rmarandfcn), '.b',
'MarkerSize', 5);
title('RMA random - FC vs BC')
xlim([-4 6])
ylim([0 25000])

%fc vs. bc
subplot(1,3,2)
hold on;
plot(gcrmarandfc(gcrmarandfcp), bcgcrmarand(gcrmarandfcp), '.r',
'MarkerSize', 5);
plot(gcrmarandfc(gcrmarandfcn), bcgcrmarand(gcrmarandfcn), '.b',
'MarkerSize', 5);
title('gcRMA random - FC vs BC')
xlim([-4 6])
ylim([0 25000])

%fc vs. bc
subplot(1,3,3)
hold on;
plot(mas5randfc(mas5randfcp), bcmas5rand(mas5randfcp), '.r',
'MarkerSize', 5);
plot(mas5randfc(mas5randfcn), bcmas5rand(mas5randfcn), '.b',
'MarkerSize', 5);
title('MAS5 random - FC vs BC')
xlim([-4 6])
ylim([0 25000])

%fc vs. dd
figure;
subplot(1,3,1)
hold on;
plot(rmarandfc(rmarandfcp), ddrmarand(rmarandfcp), '.r',
'MarkerSize', 5);
plot(rmarandfc(rmarandfcn), ddrmarand(rmarandfcn), '.b',
'MarkerSize', 5);
title('RMA random - FC vs DD')
xlim([-4 6])
ylim([0 500])

%fc vs. dd
subplot(1,3,2)
hold on;
plot(gcrmarandfc(gcrmarandfcp), ddgcrmarand(gcrmarandfcp), '.r',
'MarkerSize', 5);
plot(gcrmarandfc(gcrmarandfcn), ddgcrmarand(gcrmarandfcn), '.b',
'MarkerSize', 5);
title('gcRMA random - FC vs DD')
xlim([-4 6])
ylim([0 500])

%fc vs. dd
subplot(1,3,3)
hold on;
plot(mas5randfc(mas5randfcp), ddmas5rand(mas5randfcp), '.r',
'MarkerSize', 5);
plot(mas5randfc(mas5randfcn), ddmas5rand(mas5randfcn), '.b',
'MarkerSize', 5);
title('MAS5 random - FC vs DD')
xlim([-4 6])
ylim([0 500])
```

```
%pval vs. dd
figure;
subplot(1,3,1)
hold on;
plot(rmarandpval(rmarandfcp), ddrmarand(rmarandfcp), '.r',
'MarkerSize', 5);
plot(rmarandpval(rmarandfcn), ddrmarand(rmarandfcn), '.b',
'MarkerSize', 5);
title('RMA random - Pval vs DD (upregulated:red)')
ylim([0 500])

%pval vs. dd
subplot(1,3,2)
hold on;
plot(gcrmarandpval(gcrmarandfcp), ddgcrmarand(gcrmarandfcp), '.r',
'MarkerSize', 5);
plot(gcrmarandpval(gcrmarandfcn), ddgcrmarand(gcrmarandfcn), '.b',
'MarkerSize', 5);
title('gcRMA random - Pval vs DD (upregulated:red)')
ylim([0 500])

%pval vs. dd
subplot(1,3,3)
hold on;
plot(mas5randpval(mas5randfcp), ddmas5rand(mas5randfcp), '.r',
'MarkerSize', 5);
plot(mas5randpval(mas5randfcn), ddmas5rand(mas5randfcn), '.b',
'MarkerSize', 5);
title('MAS5 random - Pval vs DD (upregulated:red)')
ylim([0 500])


%pval vs. cc
figure;
subplot(1,3,1)
plot(rmarandpval(rmarandfcp), ccrmarand(rmarandfcp), '.r',
'MarkerSize', 5);
hold on;
plot(rmarandpval(rmarandfcn), ccrmarand(rmarandfcn), '.b',
'MarkerSize', 5);
title('RMA random - Pval vs CC (upregulated:red)')

%pval vs. cc
subplot(1,3,2)
plot(gcrmarandpval(gcrmarandfcp), ccgcrmarand(gcrmarandfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmarandpval(gcrmarandfcn), ccgcrmarand(gcrmarandfcn), '.b',
'MarkerSize', 5);
title('gcRMA random - Pval vs CC (upregulated:red)')

%pval vs. cc
subplot(1,3,3)
plot(mas5randpval(mas5randfcp), ccmas5rand(mas5randfcp), '.r',
'MarkerSize', 5);
hold on;
plot(mas5randpval(mas5randfcn), ccmas5rand(mas5randfcn), '.b',
'MarkerSize', 5);
title('MAS5 random - Pval vs CC (upregulated:red)')
```

```
%pval vs. bc
figure;
subplot(1,3,1)
plot(rmarandpval(rmarandfcp), bcrmarand(rmarandfcp), '.r',
'MarkerSize', 5);
hold on;
plot(rmarandpval(rmarandfcn), bcrmarand(rmarandfcn), '.b',
'MarkerSize', 5);
title('RMA random - Pval vs BC (upregulated:red)')
ylim([0 25000])

%pval vs. bc
subplot(1,3,2)
plot(gcrmarandpval(gcrmarandfcp), bcgcrmarand(gcrmarandfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmarandpval(gcrmarandfcn), bcgcrmarand(gcrmarandfcn), '.b',
'MarkerSize', 5);
title('gcRMA random - Pval vs BC (upregulated:red)')
ylim([0 25000])

%pval vs. bc
subplot(1,3,3)
plot(mas5randpval(mas5randfcp), bcmas5rand(mas5randfcp), '.r',
'MarkerSize', 5);
hold on;
plot(mas5randpval(mas5randfcn), bcmas5rand(mas5randfcn), '.b',
'MarkerSize', 5);
title('MAS5 random - Pval vs BC (upregulated:red)')
ylim([0 25000])



%SCATTER PLOTS
%cc
figure;
subplot(1,3,1)
scatter(ccrmarand, ccgcrmarand, '.b');
xlim([0 1])
ylim([0 1])
title('Clustering Coefficient - random - RMA vs gcRMA');
subplot(1,3,2)
scatter(ccrmarand, ccmas5rand, '.b');
xlim([0 1])
ylim([0 1])
title('Clustering Coefficient - random - RMA vs MAS5');
subplot(1,3,3)
scatter(ccgcrmarand, ccmas5rand, '.b');
xlim([0 1])
ylim([0 1])
title('Clustering Coefficient - random - gcRMA vs MAS5');


%bc
figure;
subplot(1,3,1)
scatter(bcrmarand, bcgcrmarand, '.b');
xlim([0 25000])
ylim([0 25000])
```

```
title('Betweenness Centrality - random - RMA vs gcRMA');
subplot(1,3,2)
scatter(bcrmarand, bcmas5rand, '.b');
xlim([0 25000])
ylim([0 25000])
title('Betweenness Centrality - random - RMA vs MAS5');
subplot(1,3,3)
scatter(bcgcrmarand, bcmas5rand, '.b');
xlim([0 25000])
ylim([0 25000])
title('Betweenness Centrality - random - gcRMA vs MAS5');

%dd
figure;
subplot(1,3,1)
scatter(ddrmarand, ddgcrmarand, '.b');
xlim([0 500])
ylim([0 500])
title('Degree Distribution - random - RMA vs gcRMA');
subplot(1,3,2)
scatter(ddrmarand, ddmas5rand, '.b');
xlim([0 500])
ylim([0 500])
title('Degree Distribution - random - RMA vs MAS5');
subplot(1,3,3)
scatter(ddgcrmarand, ddmas5rand, '.b');
xlim([0 500])
ylim([0 500])
title('Degree Distribution - random - gcRMA vs MAS5');
```

Plotting intersection data:

```
%calculating network topology measures for intersection data

rmaintpval = rmapvalues(intersectprobesindex);
rmaintfc = mean(intersectrmadata(:,hypoxiaindex)') -
mean(intersectrmadata(:,normoxiaindex)');

gcrmaintpval = gcrmapvalues(intersectprobesindex);
gcrmaintfc = mean(intersectgcrmadata(:,hypoxiaindex)') -
mean(intersectgcrmadata(:,normoxiaindex)');

mas5intpval = mas5pvalues(intersectprobesindex);
mas5intfc = mean(intersectmas5data(:,hypoxiaindex)') -
mean(intersectmas5data(:,normoxiaindex)');

%fc vs. cc
figure;
subplot(1,3,1);
rmaintfcp = rmaintfc > 0;
rmaintfcn = rmaintfc < 0;
plot(rmaintfc(rmaintfcp), ccintrma(rmaintfcp), '.r', 'MarkerSize',
5);
hold on;
plot(rmaintfc(rmaintfcn), ccintrma(rmaintfcn), '.b', 'MarkerSize',
5);
title('RMA intersection - FC vs CC')
ylim([0.55 1])
xlim([-5 10])
```

```
%fc vs. cc
subplot(1,3,2);
gcrmaintfcp = gcrmaintfc > 0;
gcrmaintfcn = gcrmaintfc < 0;
plot(gcrmaintfc(gcrmaintfcp), ccintgcrma(gcrmaintfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmaintfc(gcrmaintfcn), ccintgcrma(gcrmaintfcn), '.b',
'MarkerSize', 5);
title('gcRMA intersection - FC vs CC')
ylim([0.55 1])
xlim([-5 10])

%fc vs. cc
subplot(1,3,3);
mas5intfcp = mas5intfc > 0;
mas5intfcn = mas5intfc < 0;
plot(mas5intfc(mas5intfcp), ccintmas5(mas5intfcp), '.r',
'MarkerSize', 5);
hold on;
plot(mas5intfc(mas5intfcn), ccintmas5(mas5intfcn), '.b',
'MarkerSize', 5);
title('MAS5 intersection - FC vs CC')
ylim([0.55 1])
xlim([-5 10])


%fc vs. bc
figure;
subplot(1,3,1);
plot(rmaintfc(rmaintfcp), bcintrma(rmaintfcp), '.r', 'MarkerSize',
5);
hold on;
plot(rmaintfc(rmaintfcn), bcintrma(rmaintfcn), '.b', 'MarkerSize',
5);
title('RMA intersection - FC vs BC')
ylim([0 1800])
xlim([-10 10])

%fc vs. bc
subplot(1,3,2);
plot(gcrmaintfc(gcrmaintfcp), bcintgcrma(gcrmaintfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmaintfc(gcrmaintfcn), bcintgcrma(gcrmaintfcn), '.b',
'MarkerSize', 5);
title('gcRMA intersection - FC vs BC')
ylim([0 1800])
xlim([-10 10])

%fc vs. bc
subplot(1,3,3);
plot(mas5intfc(mas5intfcp), bcintmas5(mas5intfcp), '.r',
'MarkerSize', 5);
hold on;
plot(mas5intfc(mas5intfcn), bcintmas5(mas5intfcn), '.b',
'MarkerSize', 5);
title('MAS5 intersection - FC vs BC')
ylim([0 1800])
xlim([-10 10])
```

```
%fc vs. dd
figure;
subplot(1,3,1)
plot(rmaintfc(rmaintfcp), ddintrma(rmaintfcp), '.r', 'MarkerSize',
5);
hold on;
plot(rmaintfc(rmaintfcn), ddintrma(rmaintfcn), '.b', 'MarkerSize',
5);
title('RMA intersection - FC vs DD')
xlim([-5 10])
ylim([0 1200])

%fc vs. dd
subplot(1,3,2)
plot(gcrmaintfc(gcrmaintfcp), ddintgcrma(gcrmaintfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmaintfc(gcrmaintfcn), ddintgcrma(gcrmaintfcn), '.b',
'MarkerSize', 5);
title('gcRMA intersection - FC vs DD')
xlim([-5 10])
ylim([0 1200])

%fc vs. dd
subplot(1,3,3)
plot(mas5intfc(mas5intfcp), ddintmas5(mas5intfcp), '.r',
'MarkerSize', 5);
hold on;
plot(mas5intfc(mas5intfcn), ddintmas5(mas5intfcn), '.b',
'MarkerSize', 5);
title('MAS5 intersection - FC vs DD')
xlim([-5 10])
ylim([0 1200])



%pval vs. dd
figure;
subplot(1,3,1)
plot(rmaintpval(rmaintfcp), ddintrma(rmaintfcp), '.r', 'MarkerSize',
5);
hold on;
plot(rmaintpval(rmaintfcn), ddintrma(rmaintfcn), '.b', 'MarkerSize',
5);
title('RMA intersection - Pval vs DD (upregulated:red)')
ylim([1 1200])

%pval vs. dd
subplot(1,3,2)
plot(gcrmaintpval(gcrmaintfcp), ddintgcrma(gcrmaintfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmaintpval(gcrmaintfcn), ddintgcrma(gcrmaintfcn), '.b',
'MarkerSize', 5);
title('gcRMA intersection - Pval vs DD (upregulated:red)')
ylim([1 1200])

%pval vs. dd
subplot(1,3,3)
plot(mas5intpval(mas5intfcp), ddintmas5(mas5intfcp), '.r',
'MarkerSize', 5);
```

```
hold on;
plot(mas5intpval(mas5intfcn), ddintmas5(mas5intfcn), '.b',
'MarkerSize', 5);
title('MAS5 intersection - Pval vs DD (upregulated:red)')
ylim([1 1200])


%pval vs. cc
figure;
subplot(1,3,1)
plot(rmaintpval(rmaintfcp), ccintrma(rmaintfcp), '.r', 'MarkerSize',
5);
hold on;
plot(rmaintpval(rmaintfcn), ccintrma(rmaintfcn), '.b', 'MarkerSize',
5);
title('RMA intersection - Pval vs CC (upregulated:red)')
ylim([0.55 1])

%pval vs. cc
subplot(1,3,2)
plot(gcrmaintpval(gcrmaintfcp), ccintgcrma(gcrmaintfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmaintpval(gcrmaintfcn), ccintgcrma(gcrmaintfcn), '.b',
'MarkerSize', 5);
title('gcRMA intersection - Pval vs CC (upregulated:red)')
ylim([0.55 1])

%pval vs. cc
subplot(1,3,3)
plot(mas5intpval(mas5intfcp), ccintmas5(mas5intfcp), '.r',
'MarkerSize', 5);
hold on;
plot(mas5intpval(mas5intfcn), ccintmas5(mas5intfcn), '.b',
'MarkerSize', 5);
title('MAS5 intersection - Pval vs CC (upregulated:red)')
ylim([0.55 1])


%pval vs. bc
figure;
subplot(1,3,1)
plot(rmaintpval(rmaintfcp), bcintrma(rmaintfcp), '.r', 'MarkerSize',
5);
hold on;
plot(rmaintpval(rmaintfcn), bcintrma(rmaintfcn), '.b', 'MarkerSize',
5);
title('RMA intersection - Pval vs bc (upregulated:red)')
ylim([0 1800])

%pval vs. bc
subplot(1,3,2)
plot(gcrmaintpval(gcrmaintfcp), bcintgcrma(gcrmaintfcp), '.r',
'MarkerSize', 5);
hold on;
plot(gcrmaintpval(gcrmaintfcn), bcintgcrma(gcrmaintfcn), '.b',
'MarkerSize', 5);
title('gcRMA intersection - Pval vs bc (upregulated:red)')
ylim([0 1800])
```

```
%pval vs. bc
subplot(1,3,3)
plot(mas5intpval(mas5intfcp), bcintmas5(mas5intfcp), '.r',
'MarkerSize', 5);
hold on;
plot(mas5intpval(mas5intfcn), bcintmas5(mas5intfcn), '.b',
'MarkerSize', 5);
title('MAS5 intersection - Pval vs bc (upregulated:red)')
ylim([0 1800])



%SCATTER PLOTS
%cc
figure;
subplot(1,3,1)
scatter(ccintrma(rmaintfcn), ccintgcrma(gcrmaintfcn), '.b');
hold on;
scatter(ccintrma(rmaintfcp), ccintgcrma(gcrmaintfcp), '.r');
title('Clustering Coefficient - intersection - RMA vs gcRMA');
xlim([0.5 1])
ylim([0.5 1])

subplot(1,3,2)
scatter(ccintrma(rmaintfcn), ccintmas5(mas5intfcn), '.b');
hold on;
scatter(ccintrma(rmaintfcp), ccintmas5(mas5intfcp), '.r');
title('Clustering Coefficient - intersection - RMA vs MAS5');
xlim([0.5 1])
ylim([0.5 1])

subplot(1,3,3)
scatter(ccintgcrma(gcrmaintfcn), ccintmas5(mas5intfcn), '.b');
hold on;
scatter(ccintgcrma(gcrmaintfcp), ccintmas5(mas5intfcp), '.r');
title('Clustering Coefficient - intersection - gcRMA vs MAS5');
xlim([0.5 1])
ylim([0.5 1])


%bc
figure;
subplot(1,3,1)
scatter(bcintrma(rmaintfcn), bcintgcrma(gcrmaintfcn), '.b');
hold on;
scatter(bcintrma(rmaintfcp), bcintgcrma(gcrmaintfcp), '.r');
title('Betweenness Centrality - intersection - RMA vs gcRMA');
xlim([0 1800])
ylim([0 1800])

subplot(1,3,2)
scatter(bcintrma(rmaintfcn), bcintmas5(mas5intfcn), '.b');
hold on;
scatter(bcintrma(rmaintfcp), bcintmas5(mas5intfcp), '.r');
title('Betweenness Centrality - intersection - RMA vs MAS5');
xlim([0 1800])
ylim([0 1800])

subplot(1,3,3)
scatter(bcintgcrma(gcrmaintfcn), bcintmas5(mas5intfcn), '.b');
hold on;
```

```
scatter(bcintgcrma(gcrmaintfcp), bcintmas5(mas5intfcp), '.r');
title('Betweenness Centrality - intersection - gcRMA vs MAS5');
xlim([0 1800])
ylim([0 1800])

%dd
figure;
subplot(1,3,1)
scatter(ddintrma(rmaintfcn), ddintgcrma(gcrmaintfcn), '.b');
hold on;
scatter(ddintrma(rmaintfcp), ddintgcrma(gcrmaintfcp), '.r');
title('Degree Distribution - intersection - RMA vs gcRMA');
xlim([0 1200])
ylim([0 1200])

subplot(1,3,2)
scatter(ddintrma(rmaintfcn), ddintmas5(mas5intfcn), '.b');
hold on;
scatter(ddintrma(rmaintfcp), ddintmas5(mas5intfcp), '.r');
title('Degree Distribution - intersection - RMA vs MAS5');
xlim([0 1200])
ylim([0 1200])

subplot(1,3,3)
scatter(ddintgcrma(gcrmaintfcn), ddintmas5(mas5intfcn), '.b');
hold on;
scatter(ddintgcrma(gcrmaintfcp), ddintmas5(mas5intfcp), '.r');
title('Degree Distribution - intersection - gcRMA vs MAS5');
xlim([0 1200])
ylim([0 1200])
```

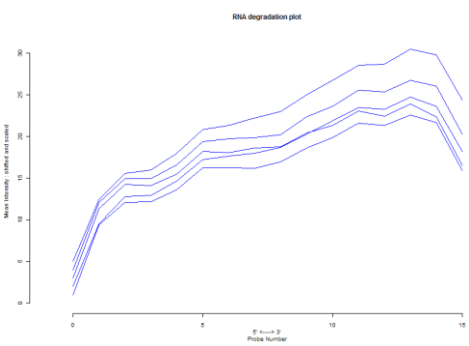
Plotting both random and intersection data:

```
%fc vs. cc
figure;
subplot(1,3,1)
hold on;
title('RMA - FC vs CC - (y:random, b:actual)');
plot(rmarandfc, ccrmarand, '.y', 'MarkerSize', 5);
plot(rmaintfc, ccintrma, '.k', 'MarkerSize', 5);
xlim([-5 10])
ylim([0.2 1])

%fc vs. cc
subplot(1,3,2)
hold on;
title('gcRMA - FC vs CC - (y:random, b:actual)');
plot(gcrmarandfc, ccgcrmarand, '.y', 'MarkerSize', 5);
plot(gcrmaintfc, ccintgcrma, '.k', 'MarkerSize', 5);
xlim([-5 10])
ylim([0.2 1])

%fc vs. cc
subplot(1,3,3)
hold on;
title('MAS5 - FC vs CC - (y:random, b:actual)');
plot(mas5randfc, ccmas5rand, '.y', 'MarkerSize', 5);
plot(mas5intfc, ccintmas5, '.k', 'MarkerSize', 5);
xlim([-5 10])
ylim([0.2 1])
```

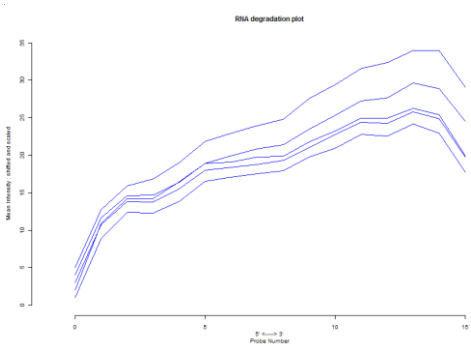
```
%fc vs. bc
figure;
subplot(1,3,1)
hold on;
title('RMA - FC vs BC - (y:random, b:actual)');
plot(rmarandfc, bcrmarand, '.y', 'MarkerSize', 5);
plot(rmaintfc, bcintrma, '.k', 'MarkerSize', 5);
xlim([-10 10])
ylim([0 25000])

%fc vs. bc
subplot(1,3,2)
hold on;
title('gcRMA - FC vs BC - (y:random, b:actual)');
plot(gcrmarandfc, bcgcrmarand, '.y', 'MarkerSize', 5);
plot(gcrmaintfc, bcintgcrma, '.k', 'MarkerSize', 5);
xlim([-10 10])
ylim([0 25000])

%fc vs. bc
subplot(1,3,3)
hold on;
title('MAS5 - FC vs BC - (y:random, b:actual)');
plot(mas5randfc, bcmas5rand, '.y', 'MarkerSize', 5);
plot(mas5intfc, bcintmas5, '.k', 'MarkerSize', 5);
xlim([-10 10])
ylim([0 25000])



%fc vs. dd
figure;
subplot(1,3,1)
hold on;
title('RMA - FC vs DD - (y:random, b:actual)');
plot(rmarandfc, ddrmarand, '.y', 'MarkerSize', 5);
plot(rmaintfc, ddintrma, '.k', 'MarkerSize', 5);
xlim([-5 10])

%fc vs. dd
subplot(1,3,2)
hold on;
title('gcRMA - FC vs DD - (y:random, b:actual)');
plot(gcrmarandfc, ddgcrmarand, '.y', 'MarkerSize', 5);
plot(gcrmaintfc, ddintgcrma, '.k', 'MarkerSize', 5);
xlim([-5 10])

%fc vs. dd
subplot(1,3,3)
hold on;
title('MAS5 - FC vs DD - (y:random, b:actual)');
plot(mas5randfc, ddmas5rand, '.y', 'MarkerSize', 5);
plot(mas5intfc, ddintmas5, '.k', 'MarkerSize', 5);
xlim([-5 10])
```

APPENDIX I

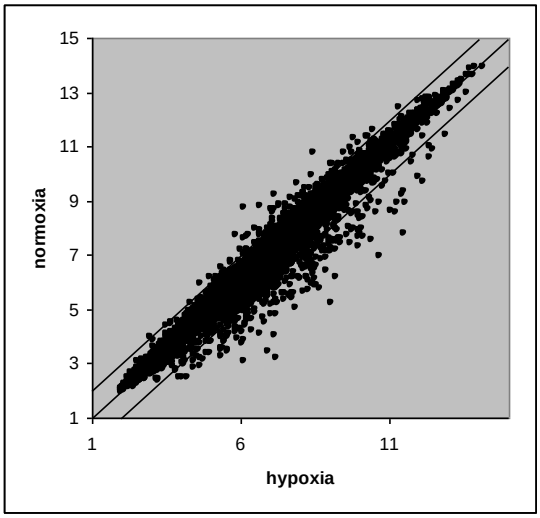
Dataset quality control measures were provided: RNA degradation plots, quality control measurements such as percent presence, background, scaling factor, and housekeeping gene 5'/3' ratios were given in the figures and tables shown below (simpleaffy; www.bioconductor.org; analyses were performed using BRB-ArrayTools developed by Dr. Richard Simon and BRB-ArrayTools Development Team.) A scatter plot between mean normoxia and hypoxia probeset intensities also was provided (BRB-ArrayTools). Two arrays exhibiting deviation from the rest of the arrays (affyPLM; bioconductor.org) were identified using NUSE parameters, and a scatter plot without these two arrays, GSM112800 and GSM112806, also was shown.



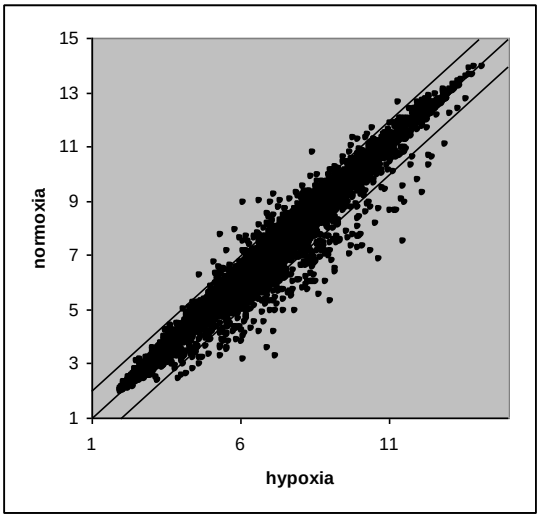
RNA degradation plots for normoxia



RNA degradation plots for hypoxia



Scatter plot of normoxia versus hypoxia



Scatter plot of normoxia versus hypoxia excluding GSM112800 and GSM112806

Table of Array Quality Measures from simpleAffy

	PercentPresent	ScaleFactor	b.Actin3.5	GapDH3.5	b.Actin3.M	GapDG3.M	AvgBackground
GSM112796.cel.present	67.420119	0.828488	0.272982	0.230186	0.968312	0.680024	32.881969
GSM112799.cel.present	66.651726	1.012586	-0.505754	0.384337	1.034074	0.752911	35.630309
GSM112801.cel.present	67.65704	0.981132	-0.30707	0.408688	0.923712	0.589918	30.327384
GSM112804.cel.present	68.399821	0.809444	0.02498	0.422888	1.049277	0.768355	35.859939
GSM112805.cel.present	72.088109	0.771758	-0.451513	0.383217	1.022426	0.642657	29.005681

	PercentPresent	ScaleFactor	b.Actin3.5	GapDH3.5	b.Actin3.M	GapDG3.M	AvgBackground
GSM112798.cel.present	67.067939	0.732321	-0.460963	0.187509	1.142593	0.696249	32.186194
GSM112800.cel.present	65.588781	0.818906	1.089244	0.316111	1.56647	0.868602	32.639049
GSM112802.cel.present	65.710444	0.883537	0.122377	0.156951	0.957879	0.774765	33.62756
GSM112803.cel.present	64.545047	0.996928	0.096361	0.313394	1.158403	0.768255	34.338878
GSM112806.cel.present	69.501185	0.864254	0.637525	0.610015	1.739151	0.961237	34.188487

Table of NUSE statististics from the AffyPLM package

```
> NUSE(Pset, type = "stats")
       GSM112796.CEL GSM112798.CEL GSM112799.CEL GSM112800.CEL GSM112801.CEL
median    0.99777075    0.99314020    0.99615531    1.02209952    0.99993841
IQR       0.02303366    0.02139885    0.02148714    0.04458799    0.02612786
       GSM112802.CEL GSM112803.CEL GSM112804.CEL GSM112805.CEL GSM112806.CEL
median    0.99670464    0.99809864    0.99303578    1.00361457    1.01986876
IQR       0.02123813    0.02233189    0.02123398    0.02818807    0.03965814
```