

DESIGN AND ANALYSIS OF LARGE SCALE GENE EXPRESSION
EXPERIMENTS AND THE APPLICATION TO ANGIOGENESIS AND BLOOD
VESSEL MATURATION

by

Kevin Anthony Greer

A Dissertation Submitted to the Faculty of the
GRADUATE INTERDISCIPLINARY PROGRAM IN BIOMEDICAL ENGINEERING
In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF PHILOSOPHY
In the Graduate College
THE UNIVERSITY OF ARIZONA

2006

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Kevin Anthony Greer entitled Design and analysis of large scale gene expression experiments and the application to angiogenesis and blood vessel maturation and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: 12/9/2005
James Hoying

_____ Date: 12/9/2005
Bob Collier

_____ Date: 12/9/2005
Scott Klewer

_____ Date: 12/9/2005
Stuart Williams

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 12/9/2005
Dissertation Director: James B. Hoying

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Kevin Anthony Greer

ACKNOWLEDGEMENTS

I would first like to thank my dissertation committee members, Jay Hoying, Stu Williams, Bob Collier, and Scott Klewer.

In particular I would like to thank my advisor and mentor Jay Hoying to whom I am eternally grateful.

I would also like to thank all of the people who I have worked with over the years, of which there are too many to list.

Most importantly, I would like to thank my wife Linda and parents Bev and Dwaine, without whom this dissertation would not have been possible.

DEDICATIONS

To my family and friends.

TABLE OF CONTENTS

LIST OF FIGURES	7
LIST OF TABLES	9
ABSTRACT	10
1 INTRODUCTION TO GENOMICS	12
1.1 Comparative Genomics	15
1.2 Quantitative Trait Loci Mapping	19
1.3 Functional Genomics	23
1.4 Specific Aims	26
2 CARMA: COMPUTATIONAL ANALYSIS OF REPLICATED MEASURES FOR ARRAYS	30
2.1 Introduction	30
2.2 Implementation	43
2.3 Results and Discussion	46
2.4 Conclusion	64
3 AGGLOMERATIVE HIERARCHICAL CLUSTERING	70
3.1 Introduction	70
3.2 Materials and Methods	82
3.3 Results	93
3.4 Discussion	102
4 GENE EXPRESSION ANALYSIS ON AN IN-VIVO MODEL OF ANGIOGENESIS AND BLOOD VESSEL MATURATION	105
4.1 Introduction	105
4.2 Results	119
4.3 Discussion	142
5 CONCLUSION	148
APPENDICES	158
APPENDIX A DATA FLOW DIAGRAM FOR CARMA	158
APPENDIX B AGGLOMERATIVE HIERARCHICAL CLUSTERS FOR MOUSE MICROVESSEL EXPERIMENT	160
APPENDIX C DETAILED METHODS FOR MOUSE MICROVESSEL EXPERIMENT	182
APPENDIX D GLOSSARY OF TERMS	187
REFERENCES	191

LIST OF FIGURES

Figure 1.1 Comparison of DNA sequence between species for VEGFA	18
Figure 1.2 Quantitative trait loci mapping for rat chromosome 1	22
Figure 2.1 Overview of microarray production, hybridization, and scanning.....	33
Figure 2.2 Screen capture of a microarray image during spot finding	34
Figure 2.3 Microarray hybridization schemes	37
Figure 2.4 Input form for generating simulated microarray for use with ANOVA.....	45
Figure 2.5 Comparison of \log_2 and \ln transformations.....	52
Figure 2.6 Comparison of pre-normalized and post-normalized data	55
Figure 2.7 Normalization of a two-channel hybridization.....	58
Figure 2.8 Results of ANOVA for the <i>Mus Musculus Plat</i> gene	63
Figure 3.1 Agglomerative hierarchical clustering tree	73
Figure 3.2 One-dimensional self organizing map.....	75
Figure 3.3 Input form for generating simulated microarray data for clustering	84
Figure 3.4 Simulated microarray datasets.....	86
Figure 3.5 Mean centering and unit normalization.....	87
Figure 3.6 Progression of cluster recovery and cluster number for all algorithms.....	97
Figure 4.1 Vascular endothelial growth-factor receptor intracellular signaling	111
Figure 4.2 Images of microvessel implants explanted at days 7, 14, 21, 28.....	121
Figure 4.3 Mouse microvessel explants stained for smooth muscle actin.....	122
Figure 4.4 Intravital images of blood flow in microvascular implants.....	124
Figure 4.5 Images of implants after intravascular injection of rhodamine dextran	125

Figure 4.6 Cell proliferation at 3, 7, 14, 21, and 28 days post implantation.....	127
Figure 4.7 Hybridization scheme for mouse microvessel experiment.....	129
Figure 4.8 Mouse microvessel gene expression clusters for timecourse experiment	131
Figure 4.9 Principal components of mouse microvessel gene expression data	136
Figure 4.10 Project of the original time points onto first 3 principal components	138

LIST OF TABLES

Table 3.1 Comparison of cluster recovery	96
Table 4.1 Endogenous inhibitors of angiogenesis	108
Table 4.2 Real-time PCR measurements of gene expression for select genes	133
Table 4.3 Principal components of mouse microvessel gene expression data	135
Table 4.4 Overrepresented Gene Ontology categories	140

ABSTRACT

The objective of this dissertation was to develop an experimental approach and supporting software for performing and interpreting the results of microarray-based experiments, as well as apply this approach to an experimental model of angiogenesis and blood vessel development. When this project was initiated microarray technology was in its infancy and the standard experimental design was to hybridize two samples against each other and report intensity ratios that were greater than two-fold. In order to study the changes in gene expression that occur over the course of the vascularization process, it became clear that a new approach to microarray experimental design and analysis was required. It was also clear that most researchers were ill-equipped to process and interpret the tens of thousands of data points generated by microarray experiments. To address these needs, a software package called CARMA (Computational Analysis of Replicated Measurements for Arrays) was developed to perform an analysis of variance (ANOVA) on microarray experiments that incorporate replicated measurements. Utilizing replicated measurement-based designs makes it possible to incorporate multiple samples into the experimental design and calculate both the magnitude and the statistical significance of the differences in gene expression between samples. Software was also developed to implement and compare different algorithms and distance metrics for performing hierarchical clustering. Hierarchical clustering groups genes together based on the similarity of their expression profiles, and is used to reduce the complexity of a microarray dataset and identify genes that may be involved in the same or related

processes or under similar types of transcriptional control. Utilizing simulated datasets containing known clusters of genes, the ability of each algorithm/distance metric combination to recover the original clusters was evaluated. Lastly, both CARMA and hierarchical clustering were utilized to analyze changes in gene expression during the process of vascularization in an experimental model of angiogenesis and blood vessel maturation. Based on high-level patterns of gene expression and morphological measurements obtained using this model, a multi-phase model of angiogenesis-based vascularization is presented consisting of an initial angiogenic phase, followed by a maturation and network remodeling phase.

1 INTRODUCTION TO GENOMICS

The term *genomics* was first suggested by mouse geneticist T. H. Roderick in 1987 to describe the newly developing discipline of nucleic acid mapping/sequencing. Its definition was published in the first issue of the journal bearing the same name; *Genomics* volume one issue one states “The new discipline is born from a marriage of molecular and cell biology with classical genetics and is fostered by computational science. Genomics involves workers competent in constructing and interpreting various types of genomic maps and interested in learning their biologic significance. Genetic mapping and nucleic acid sequencing should be viewed as parts of the same analytic process—a process intertwined with our efforts to understand development and disease.”(McKusick & Ruddle, 1987). Over the past twenty years the field of genomics has been expanded to include: comparative genomics, bioinformatics and computational biology, functional genomics, quantitative trait loci mapping, and new genomics technologies(Boguski, 2005). Genomics, however, would never have been possible without the field of genetics.

Although relatively unknown until years after his death, Gregor Mendel first described the basic rules of genetic inheritance in 1866(Mendel, 1866). Through controlled cross-breeding of common pea plants over numerous generations, Mendel realized that certain traits are passed on to offspring without any blending of parent characteristics. This discovery, although common knowledge today, was contrary to the most accepted theory of the day in which inherited traits were thought to be blended from generation to generation. Based on his cross-breeding experiments Mendel concluded

that each plant contained two “factors” (now termed alleles) with only one allele passing from parent to progeny for each trait. Mendel also deduced that the pairs of alleles for each trait are inherited independently of each other(O'Neil D, 2005). It would take, however, another 40 years for the field of genetics to be born.

William Bateson coined the term *genetics* at the first International Congress of Botany in 1906 by stating “a new and well developed branch of Physiology has been created. To this study we may give the title Genetics”. However, the word *gene* was not defined until three years later by Wilhelm Johannsen stating “The word ‘gene’ ... expresses only the evident fact that, in any case, many characteristics of the organism are specified in the gametes by means of special conditions, foundations, and determiners which are present in unique, separate, and thereby independent ways ...”. He later added “The ‘gene’ is nothing but a very applicable little word, easily combined with others, and hence it may be useful as an expression for the ‘unit factors,’ ‘elements’ or ‘allelomorphs’ in the gametes, demonstrated by modern Mendelian researches ... As to the nature of the ‘genes,’ it is as yet of no value to propose any hypothesis; but that the notion of the ‘gene’ covers a reality is evident in Mendelism.”(Keller, 2000)

Although its function was not yet known, deoxyribonucleic acid (DNA) was first discovered in 1869 by Johann Miescher, and chromosomes were discovered in 1882 by Walther Flemming. In 1888, Theodore Boveri established that chromosomes remain organized through the process of cell division, and that both the egg and sperm each contribute that same number of chromosomes during fertilization. After the rediscovery of Mendel’s findings in 1900, Boveri, along with Walter Sutton (in 1902) proposed that

chromosomes contained the material of heredity, which was later proven through breeding of the common fruit fly, *Drosophila melanogaster*, by Thomas Morgan in 1915(Morgan *et al.*, 1915). Morgan also defined the concepts of homologous recombination and genetic linkage, and together with Alfred Sturtevant created the first gene map in 1913(A.H.Sturtevant, 1913).

The most publicized discovery in genetics was the molecular structure of DNA and the hypothesis of its possible mechanism for copying genetic information made by Francis Crick and James Watson in 1953(Watson & Crick, 1953). This work explained the possible mechanism for DNA-based inheritance that was demonstrated the year before by Alfred Hershey and Martha Chase(Hershey & Chase, 1952). In 1957, Francis Crick proposed the *Central Dogma* hypothesis, where genetic information is transcribed from DNA to RNA and then translated to protein, but never in the reverse direction. Crick also proposed that DNA encoded for proteins based on a “triplet” code where three nucleotides coded each protein(Crick, 1958). Amazingly, Crick’s hypotheses predated the discovery of mRNA(JACOB & Monod, 1961) and the “genetic code”(Nirenberg *et al.*, 1962) by over three years.

The initial groundwork for the field of genomics was laid in the 1970’s. The first restriction enzyme(Smith & Wilcox, 1970), recombinant DNA molecule(Jackson *et al.*, 1972), insertion of cloned DNA(Cohen *et al.*, 1973), and DNA sequencing techniques(Maxam & Gilbert, 1977)(Sanger *et al.*, 1977) were all published in the 1970’s, along with the use of restriction fragment length polymorphisms (RFLPs). The final key invention that ushered in the era of genomics research was the development of a

simple method for amplifying DNA in the 1980's. This technique known as polymerase chain reaction (PCR)(Mullis *et al.*, 1986)(Saiki *et al.*, 1988) could be used to produce unlimited quantities of genetic sequences, and can amplify both genomic DNA or RNA, after it has been reverse transcribed into cDNA. The development of automation and high-throughput(Smith *et al.*, 1986) techniques for DNA sequencing and the production of cDNA libraries, which could then be sequenced to generate clone-sets of expressed sequence tags(Adams *et al.*, 1991), finally provided the mechanisms to perform research at the genomic scale in the early 1990's. Since then scientific advances such as microarray technology(Schena *et al.*, 1995) and quantitative real-time PCR analysis(Heid *et al.*, 1996)(Chiang *et al.*, 1996) have made functional genomic research a reality for many researchers, and large genome sequencing projects such as the human genome project(Collins *et al.*, 2003) have provided detailed genetic information of both humans and many experimental organisms(Genomics Proteomics Bioinformatics, 2004)(J.Craig Venter Institute, 2005)(Keller, 2000).

1.1 Comparative Genomics

The development of rapid sequencing technologies and computer algorithms to align and interpret genetic sequences has not only improved many of the more traditional methods of genomic/genetic research, but it has also enabled a completely new field of research. "Comparative Genomics" is the large-scale process of aligning sequences from two or more organisms to identify conserved or related sequences (Figure 1.1). Not only does this type of comparison have obvious utility in the field of evolution, but it has had a far-reaching effect on many areas of scientific research. Comparative genomics is based

upon the underlying assumption that functionally relevant sequences are conserved during evolution. As expected, comparisons of known protein coding regions (exons) between species have demonstrated that these regions are preferentially conserved (Waterston *et al.*, 2002). Utilizing this approach coding regions in one species can also be used to improve the prediction of unidentified genes in other species as well as identify undiscovered members of existing gene families (Pennacchio *et al.*, 2001). One of the most revolutionary discoveries to come out of the field of comparative genomics is the percentage of DNA that is under (purifying) selection. A comparison of the mouse and human genomes estimates that approximately 5% of the mammalian genome is under selection owing to functional constraints. Considering that 1.5% of the mammalian genome is estimated to be coding regions, the remaining 3.5% must be conserved due to non-coding mechanisms (Pennacchio, 2003). Comparisons of these non-coding regions have yielded the expected areas of conservation within promoter regions (Landry *et al.*, 2005) as well as within introns (Wardrop & Brown, 2005) and intergenic regions (Loots *et al.*, 2000).

Comparative genomics also facilitates the process of identifying positively selected regions of DNA. These regions accumulate changes more frequently than other areas of the genome and are therefore difficult to detect using other methods. Some researchers however, are particularly interested in these regions as they often represent adaptive mechanisms that differentiate between species. Examples of positively selected regions are genes encoding proteins involved in defense against pathogens, such as human histo-compatibility determinants or genes necessary for adaptation to a new

environment(Hughes & Nei, 1988), such as lysozyme in langur monkeys(Messier & Stewart, 1997)(Miller *et al.*, 2004). Comparative genomics has also confirmed that rodents accumulate nucleotide substitutions at twice the rate of humans, and that the mammalian genome has evolved non-uniformly across the genome. Not only are these two findings important from an evolution standpoint, but they must be taken into considering when identify conserved sequences with potential functionality as they affect the evolutionary distance between organisms(Pennacchio, 2003).

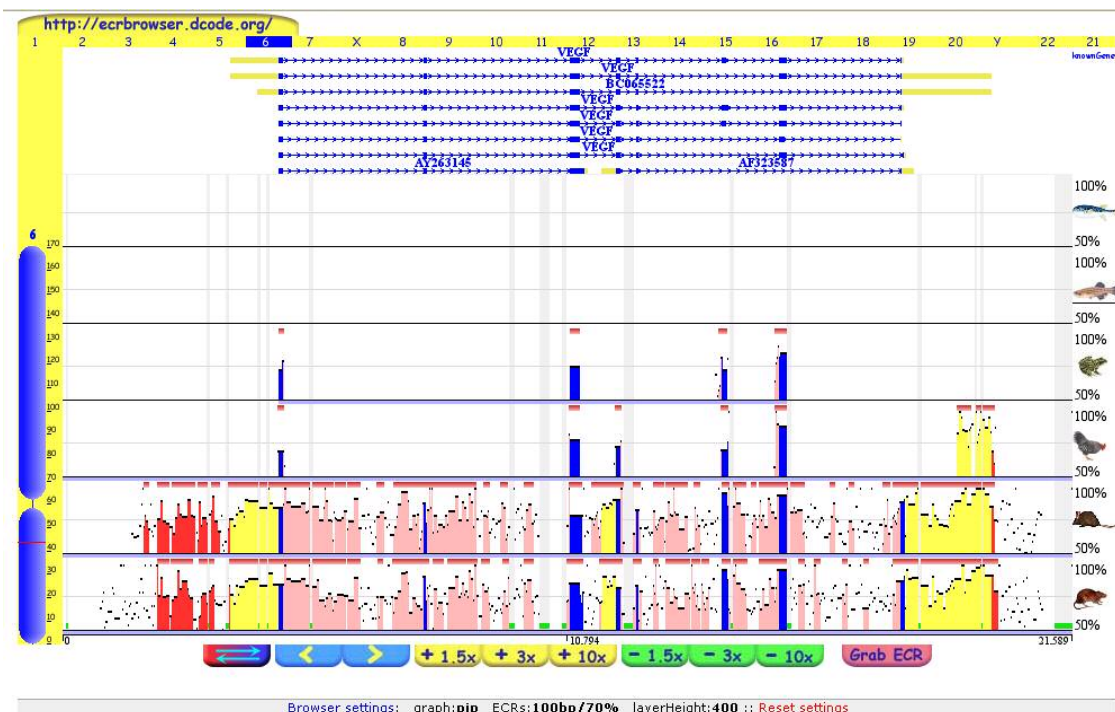


Figure 1.1 Comparison of DNA sequence between species for VEGFA

The output of a comparison between human, mouse, rat, chicken, frog, zebra fish, and fugu fish for the genomic region around the VEGFA gene. The top panel shows various mRNA sequences composed of different combinations of the 8 exons (shown as blue boxes) found in human VEGFA. The bottom six panels (one for each species) displays the percent conserved sequence (as represented by the height of the vertical line) after being aligned to the human genome. The vertical lines in each of the panels are colored as follows: blue = coding exon, yellow = untranslated, red = intergenic, pink = intron. A red box above the lines indicates an evolutionarily conserved region, which for this case has been defined as regions at least 100 bp long and at least 70% homology with the human genome (Ovcharenko *et al.*, 2004).

1.2 *Quantitative Trait Loci Mapping*

Long before the discovery of the structure and function of DNA, researchers were creating linkage maps of genetic markers for plants based on gross morphological characters such as dwarfism, albinism, and altered leaf structure (Tanksley *et al.*, 1989). These original maps attempted to order simple Mendelian inheritable traits based on their frequency of inheritance with other heritable traits. Heritable traits that are controlled by genes on separate chromosomes demonstrate no association, however genes that are on the same chromosome are “linked”, and are inherited together more frequently than not. This frequency of co-inheritance is directly correlated with homologous recombination and is a measure of how frequently recombination events occur between the two genes. These calculations of recombination events are then converted into genetic distance, measured in centimorgans, and compiled into a genetic map (A.H. Sturtevant, 1913).

Quantitative genetics is the study of inheritance associated with continuous (i.e. non-discrete) traits. The genetic contributions to these continuous traits are often influenced by environmental factors and/or multiple genes and their interactions (epistatic effects), which both complicate the process of identifying the underlying genetic cause. However it is often these quantitative traits, such as hypertension, crop yield, salt tolerance, or milk production that are of the most interest. Karl Sax first demonstrated the use of genetic markers to improve selective breeding when he statistically associated the size and color of the beans of *Phaseolus vulgaris* (Sax, 1923). Although Sax did not attempt to genetically map the locations of any of the genes for either size or weight, this publication is often cited as the first illustration of a quantitative trait locus (QTL).

Quantitative trait loci are regions of the genome that are associated with continuous traits, which when associated with a genetic map is called quantitative trait loci mapping (Figure 1.2). These original genetic maps were based on phenotypic traits, however they were rapidly replaced with molecular markers as these techniques became available. These first molecular markers, such as blood type (Renwick & Lawler, 1955) and isozymes (Hunter & Markert, 1957), were not DNA-based, however it was not until the advent of DNA markers that it was possible to create a complete human genetic linkage map (Donis-Keller *et al.*, 1987). DNA markers take advantage of the multitude of genetic polymorphisms that have accumulated in the human population during evolution. Most of these genetic differences are “silent” mutations, in that they don’t affect protein expression or function, and have no detectable phenotype.

While some DNA markers existed before the genomic era, the development of high-throughput molecular techniques and advances in computational hardware and software that has occurred over the last two decades has revolutionized the field of QTL mapping. The first major advancement in genetic markers was the advent of restriction fragment length polymorphisms (Botstein *et al.*, 1980). This process utilized bacterial restriction enzymes to cut DNA into thousands of fragments. These fragments vary in size from individual to individual based on genetic differences in restriction sites, or insertions and deletions between restriction sites. The second major advancement in genetic markers was the utilization of repetitive genetic sequences whose number varies from individual to individual. These repetitive sequences vary in size from a few nucleotides, called microsatellites (Weber & May, 1989), to a tens or hundreds of

nucleotides, called minisatellites(Jeffreys *et al.*, 1985). These markers are also referred to as variable number of tandem repeats (VNTR), because they are based on the number of times that a particular sequence is repeated. The latest advancement in genetic markers is the use of single nucleotide polymorphisms (SNPs). These markers are simply single nucleotide differences and as expected are the most prolific genetic markers estimated to differ in 1 out of every 1000 nucleotides of any two copies of a chromosome in humans(Landegren *et al.*, 1998). Most recently an international consortium charged with developing a high resolution human genetic linkage map published a 55,000 SNP map and the associated allele frequencies in African American, Asian, and European American populations(Matsuzaki *et al.*, 2004).

Genomics has also added a new dimension to QTL studies. Comparative genomics now allows researchers to contrast QTL identified in their species of interest against the corresponding genetic locations in model organisms and organisms of varying evolutionary distance. Using this approach researchers have aligned QTL with organisms that have been completely sequenced and annotated to identify possible candidate genes(Hazen *et al.*, 2003), verified QTL between species(Zimdahl *et al.*, 2002), and utilized genetically/congenically engineered experimental models to test QTL(Palijan *et al.*, 2003).

1.3 Functional Genomics

The phrase “functional genomics” seems to be used ubiquitously in scientific literature as of late, however a simple examination of each of the words in the phrase provides for a much narrower interpretation. Many authors use “functional genomics” in the context of attempting to understand the function of a gene or genes. However this interpretation ignores the second word in the phrase “genomics”. Genomics refers to an organisms complete set of genes and chromosomes. In this context, functional genomics is the pursuit of understanding gene function at the genomic scale, and is characterized by high-throughput experimental technologies coupled with statistical and computational analysis of the results(Hieter & Boguski, 1997).

The first two technologies that promised to allow gene expression monitoring at the genomic scale were published in the same issue of Science in 1995. Schena *et al.* described the development of small format high-density array of complementary DNAs

(cDNAs) that could be used to monitor the expression levels of several genes simultaneously (Schena *et al.*, 1995). These cDNAs were printed on a surface-modified glass microscope slide and could be used to measure the expression levels of their corresponding genes in two messenger RNA samples, after conversion to cDNA, using two-color fluorescent hybridization. Even though this first microarray contained only 48 genes, its parallel nature evinced its scalability to genomic gene expression monitoring. The second technology unveiled in October of 1995 was serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995). SAGE analysis utilizes standard sequencing techniques to count the number of copies of each mRNA molecule, after conversion to cDNA, based on sequencing a small portion of each transcript. Most recently a new technique for performing large scale gene expression analysis has been developed that utilized concepts from both microarrays and SAGE. Massively parallel signature sequencing utilizes 5 μm diameter microbeads and hybridization based sequencing to count the number of transcripts in a given cDNA sample (Brenner *et al.*, 2000).

Although many gene expression measurement techniques do not require a priori knowledge of genetic sequence, the complete sequence for each transcript must be deduced for subsequent analyses. Sequencing projects have completed final sequences for many organisms, however gene prediction software is unable to sufficiently predict transcript sequences (Guigo & Reese, 2005). A recent comprehensive study of the mouse transcriptome has highlighted the complexities of transcriptional regulation such as alternative promoter usage, splicing variants, and alternate polyadenylation sites (Carninci

et al., 2005). Genomic mapping of the transcriptome exposed large areas that are transcribed on either DNA strand without gaps, separated by areas devoid of transcription. In total more than 180,000 independent transcripts were identified and estimates indicated that there are ten fold more transcripts than genes. This study also reported a large number of non-coding RNAs (ncRNAs) whose promoter regions are generally more conserved than promoters on protein-coding RNAs. To further highlight the complexities of gene-expression recent evidence also suggests that antisense transcripts form normal genomic DNA can alter expression of sense mRNA(Katayama *et al.*, 2005).

Functional genomics and QTL mapping have also been combined into the field of genetical genomics. Gene expression data, for example, can be used instead of classic quantitative traits such as height or yield as a measure of a phenotype(de Koning & Haley, 2005). In this way hundreds or thousands of quantitative traits can be analyzed using the same dataset. The magnitude of these datasets requires a more rigorous statistical analysis, however combining the multitude of measurements of gene expression and sequence variation can improve both the QTL analysis and the interpretation of differences in gene expression. Depending on the recombination rate and the resolution of the genetic map, each QTL can still contain hundreds of genes. Measured expression levels for genes in the surrounding area can provide an indication of the underlying genetic cause of a QTL. Knowing the QTL profile for every gene in the functional analysis can also facilitate the construction of genetic regulatory networks. Genes with similar QTL profiles will often share some of the same regulatory elements,

and are frequently in the same genetic pathway(Alberts *et al.*, 2005). Affymetrix short nucleotide arrays provide even more detailed information about the possible causes of differences in gene expression. These arrays contain multiple probes for each mRNA. Differences in the measured fluorescence for each probe of a given sequence can point to the exact location of genetic differences, such as insertions, deletions, or SNPs based on their impact on hybridization. Affymetrix also now offers large scale genotyping microarrays that assess over 100,000 SNPs in a single reaction(Matsuzaki *et al.*, 2004).

1.4 Specific Aims

The field of genomics has dramatically expanded the tools available to researchers for obtaining gene sequence and expression information. In particular measuring gene expression at the genomic scale provides investigators with a picture of the landscape of gene expression for any given sample. Investigating changes to this gene expression landscape caused by different treatments, or associated with different conditions, should provide insight into the genetic mechanisms underpinning various diseases, treatments, physiological processes, etc. While these recent advances in molecular techniques, high throughput equipment, and computer hardware have made it possible to perform thousands of simultaneous measurements of gene expression, they have also introduced new problems associated with data management and analysis. Therefore, the overall goal of this project is to develop new computation tools for analyzing gene expression data, and utilize those tools to identify changes in gene expression in an in-vivo model of angiogenesis, blood vessel maturation, and network remodeling. These tools will be developed specifically for two-channel microarray

experiments, however the approach and resulting software should be adaptable to other types of genomic data.

Specific Aim 1. Develop software to identify significantly differentially expressed genes in microarray datasets. Utilizing microarray technology it is now possible to measure the level of transcription for thousand of genes simultaneously. Most researchers, however, are ill-equipped to process the massive amounts of data generated from microarray experiments. In an effort to identify statistically significant changes in gene expression software will be developed to perform an analysis of variance (ANOVA) on two-channel microarray datasets. In addition, this software (named CARMA - Computational Analysis of Replicate Measures for Arrays) will also perform all of the necessary steps for calculating differential expression in a microarray data set including importing, transforming, and normalizing the raw data files. Analyzing microarray data using CARMA will require only a basic understanding of the underlying principles, and output will be provided in both an easy to interpret graphical format and a delimited text file.

Specific Aim 2. Develop software to perform agglomerative hierarchical clustering and compare the effects of different clustering algorithms and distance metrics on the clustering of simulated microarray data. Agglomerative hierarchical clustering is one of the most widely used methods of grouping genes or samples based on similarity in expression profiles. This process recursively pairs genes or samples into clusters based on their expression profiles, beginning with all genes being unclustered and ending with all genes belonging to one cluster. Agglomerative hierarchical

clustering is actually a general term that refers to a family of clustering algorithms, which each employ one or more distance metrics. In an effort to better understand the effects of applying different distance metrics and clustering methodologies to cDNA microarray data a generalized computer algorithm will be developed that implements 10 hierarchical clustering methods and 4 distance metrics. Some clustering methods are usually only implemented with specific distance metrics, however this generalized algorithm will be designed such that each clustering method will work with all 4 distance metrics. The ability of each clustering method and distance metric combination to recover known clusters within simulated microarray datasets will also be assessed.

Specific Aim 3. Identify changes in gene expression during angiogenesis, blood vessel maturation, and network remodeling. In an effort to improve our understanding of the cellular mechanisms regulating angiogenesis, blood vessel maturation, and vascular remodeling, a mouse microvessel fragment model will be utilized to study gene expression during the formation of a vascular network from small vessel fragments isolated from mouse periovarial and epididymal fat pads. Following isolation, microvessel fragments are embedded in type I collagen, which is pH neutralized and allowed to gel, and implanted subcutaneously on the hindquarters of severe combined immunodeficient (SCID) mice. During the first week after implantation these microvessel fragments undergo sprouting angiogenesis to form networks of small diameter vessels. Over the course of the following three weeks these networks remodels into a typical vascular networks consisting of inflow vessels (arteries and arteriols), capillaries, and outflow vessels (veins and venules). Total RNA will be extracted from

implants explanted on days 3, 7, 14, 21, 28, as well as freshly isolated vessels (day 0).

The total RNA will then be amplified, fluorescently labeled, and comparatively

hybridized between time-points to determine changes in gene expression during the

formation of the new vasculature. Differentially expressed genes will be identified using

the software package developed as part of specific aim 1 and clustering will be performed

based on the clustering algorithms and distance metrics evaluated as part of aim 2.

2 CARMA: COMPUTATIONAL ANALYSIS OF REPLICATED MEASURES FOR ARRAYS

2.1 *Introduction*

The field of genomics was officially inaugurated only 20 years ago, however since then advances in technology and techniques have proceeded at an exponential rate. The most widely used of these technologies is solid surface (glass or silica) microarrays. The first microarray was published in 1995 (Schena *et al.*, 1995) and only contained 48 genes, however it launched an entire industry with more than 10,000 papers in pubmed containing a reference to microarrays as of October, 2005. This first cDNA microarray was developed using amplified PCR products from an Arabidopsis cDNA library. These PCR products were printed on poly-L-lysine coated microscope slides using a robot that deposits nanoliter volume spots through surface contact using a metal print tip. Surface contact printed microarrays are still the most common type of microarray used today due to their relative ease of manufacture and low cost. In addition to cDNA arrays, medium length oligo arrays consisting of single stranded DNA sequences between 50 and 90 bp have also gained popularity due to their increased specificity over cDNA arrays, however they have the disadvantage of requiring prior knowledge of each DNA sequence on the array. Oligo microarrays have the added disadvantage of only having one strand (usually the sense strand) of the DNA available for hybridization, thus limiting the choice of protocols used for RNA amplification and cDNA strand labeling.

The other major class of microarrays in wide scale use today is short oligo arrays manufactured using photolithography based processes adapted from the semiconductor industry(Lipshutz *et al.*, 1999). While these arrays offer some advantages over contact printed (spotted) arrays, they are only available from one manufacturer (Affymetrix), require specialized equipment, and are relatively expensive. However, Affymetrix microarrays have gained in popularity because of their standardized protocols and equipment, paired mismatch probes (to assess non-specific hybridization), and high density of probes. The other major difference is that only one sample at a time is hybridized to Affymetrix arrays whereas two or more samples at a time are hybridized to spotted arrays. As the focus of this dissertation is on designing and analyzing microarray experiments using spotted arrays, the remainder of this document is specific to spotted arrays hybridized with two samples. Any subsequent reference to array or microarray should be taken to mean a spotted microarray.

A spotted microarray (Figure 2.1) is simply a glass microscope slide whose surface has been modified in order to bind DNA, upon which DNA “spots” are “printed”. This modification is usually achieved through the application of a silane based coupling agent such as aminopropyltrimethoxysilane, in which the silane groups bind the glass and the functional group (i.e. amine) binds the DNA. As noted previously, the DNA that is printed on the slides can either be double stranded DNA usually obtained through PCR reactions, or single stranded manufactured oligos. The synthesized DNA is purified and aliquoted into microtiter plates, with each well containing millions of copies of one genetic sequence, and each well containing a different sequence. Each of these

sequences is referred to as a clone or expressed sequence tag (EST). These clones are then re-arrayed from the microtiter plate onto the derivatized microscope slides using a robotic printing system in which split metal pins dip into each well, draw-up liquid through capillary action, and then deposit nanoliter amounts onto each slide through brief contact. Once all of the “spots” of DNA have been deposited on the microscope slides they are subsequently referred to as microarrays.

Microarrays are used to measure the relative quantities of DNA or RNA between two samples, referred to as the targets, through hybridization with the spots of DNA on the slide, which are called probes. This process involves labeling each sample with a different fluorescent dye, combining the samples and resuspending them in a hybridization buffer, distributing the solution over the surface of the slide, and incubating the slide in an enclosed environment at a controlled temperature. Many of the DNA or RNA molecules within the samples attach to the complementary sequence on the microarray, which is then washed to remove unbound sample, and scanned at the appropriate wavelengths to generate one image for each fluorochrome. The intensity of each spot at each wavelength provides a measure of the amount DNA or RNA with the complementary sequence in each sample. Data is extracted from the images using software that aligns a grid of circles corresponding to the printed pattern of spots in the images (Figure 2.2) and reports quantitative data about the intensity the pixels within and around each circle(Qin *et al.*, 2005).

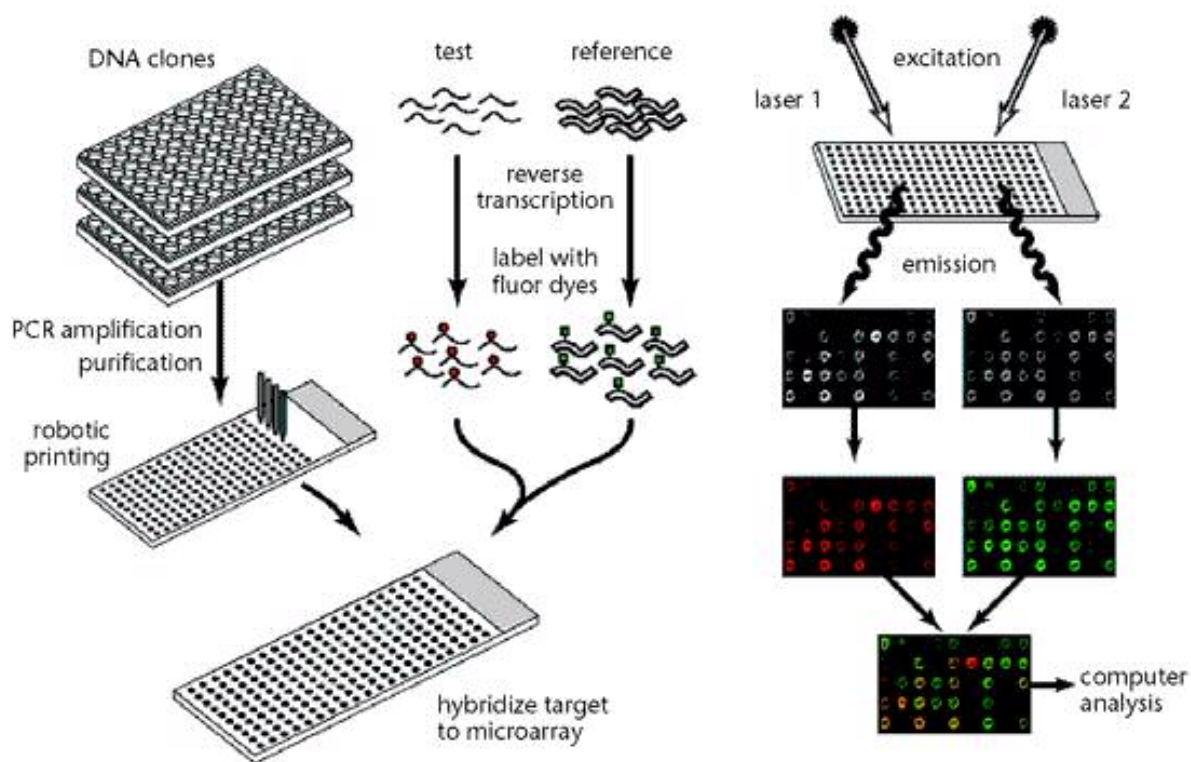


Figure 2.1 Overview of microarray production, hybridization, and scanning

(Duggan *et al.*, 1999) First the cDNA clones are printed from microtiter plates onto a glass slides that has been derivatized to bind DNA. The slides with DNA bound to their surface are subsequently referred to as microarrays. Two RNA samples are then reverse transcribed, during which time either a fluorescently labeled or chemically modified nucleotide is incorporated into the cDNA. If a chemically modified nucleotide is incorporated then fluorochromes are attached to the modified nucleotides. The fluorescently labeled cDNA is then hybridized to a microarray. The unbound cDNA is washed off of the microarray, the array is dried, and then it is scanned using a fluorescence scanner. One image is generated for each of the two wavelengths appropriate for the fluorochromes. The two images are then analyzed using image processing software in order to quantitate the levels of fluorescence for each spot and its surrounding area.

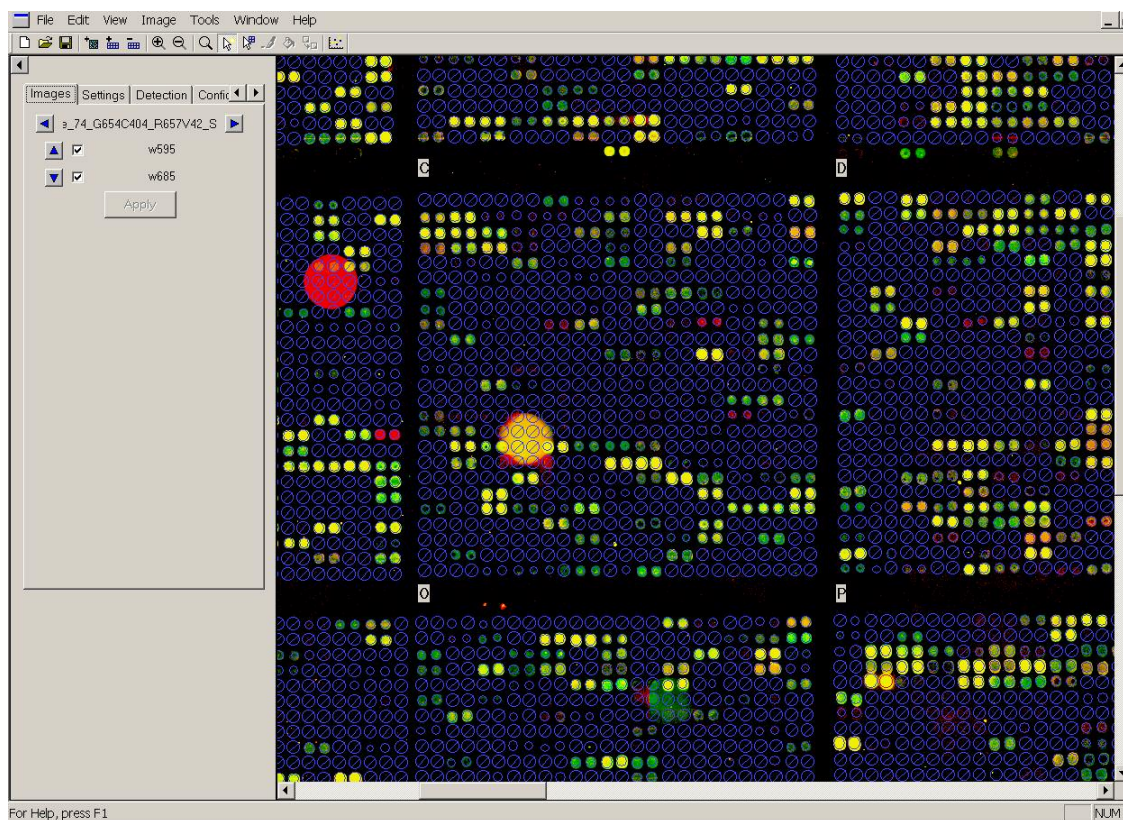


Figure 2.2 Screen capture of a microarray image during spot finding

A grid of circles is defined with the same layout as the printed microarray. The spot finding software automatically aligns the grid based on the pattern of spots within the images. Spot diameter is also automatically adjusted for each spot based on the minimum diameter of the spot. The alignment of the grid or individual circles within the grid can be manually adjusted to account for incorrect alignment. Two relatively large anomalies (large circles) can be seen in the image. Spots within these anomalies will usually be flagged automatically by the software, however users can also manually flag spots to indicate an invalid measure.

The manufacture and hybridization of spotted microarrays is a multi-step process, with each step adding to the variability in the measured fluorescence. Therefore measuring the absolute quantity of each transcript in a given sample is impossible, and even comparing between hybridizations is difficult. In order to minimize the variability between measurements, samples are usually hybridized in pairs, with each sample being labeled with a different fluorochrome in order to distinguish between the samples. Paired hybridizations minimize variability by subjecting each sample to identical hybridization conditions. Furthermore paired hybridizations encourage simultaneous sample preparation, which reduced the significant variability that can be introduced by different technicians or small differences in techniques or reagents from week-to-week(Chen *et al.*, 2004).

Initial microarray experiments focused on determining differences in gene expression between two samples(DeRisi *et al.*, 1996)(DeRisi *et al.*, 1997). When comparing more than two samples the simplest experimental design is to perform hybridizations in which each sample of interest is paired with a common reference sample, used in all hybridizations (Figure 2.3a). This “reference” design has the advantage of allowing for an unlimited number of samples, accommodates situations in which all samples are not available at the outset of the experiment, and permits expansion of the experiment to include additional samples. This type of design is also known as an indirect comparison design because comparisons between the samples of interest are made indirectly through their hybridizations with the reference sample. For smaller experiments, designs that employ direct comparisons are always more efficient(Kerr &

Churchill, 2001b). In these designs, often referred to as “loop” designs, the samples of interest are hybridized against each other, rather than against a common reference sample (Figure 2b). Their superior efficiency is readily apparent as twice as many measurements are made for the samples of interest in a loop design than in a reference design.

Unfortunately there is no standard microarray experimental design that works best in all circumstances; numerous factors including availability and cost of arrays and samples, technical and biological replication, and the intent of the experiment must be considered in order to develop the optimal experimental design (Yang & Speed, 2002).

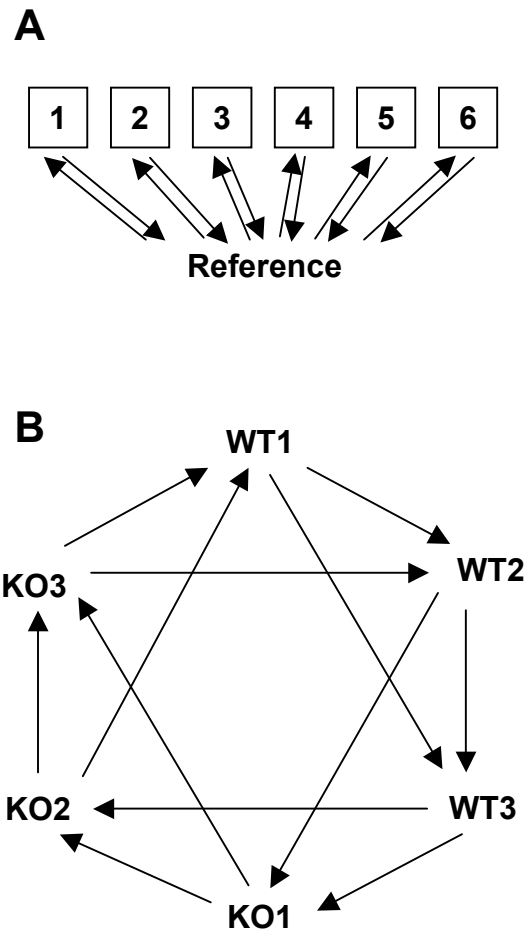


Figure 2.3 Microarray hybridization schemes

A Hybridization scheme for 6 samples where all samples are hybridized against a reference sample. **B** Interwoven loop hybridization scheme for the aquaporin-1 knockout experiment used in this study, where W1-W3 and KO1-KO3 denote each of the wild type and aquaporin-1 deficient mice respectively. Each arrow in both schemes represents one hybridization, with the tail of the arrow denoting labeling with the Alexa 546 dye and the head of the arrow denoting labeling with the Alexa 647 dye.

One of the most critical considerations in microarray experimental design is replication. There are four levels of replication possible in a microarray experiment. The first type of replication occurs when multiple spots are printed for each of the clones during the microarray fabrication process. The second type of replication takes place when multiple clones representing the same transcript are included on the microarray. These two types of replication must be taken into account during data analysis, however they have a minimal impact on experimental design. The third type of replication is called technical replication and refers to performing multiple hybridizations with the same sample. In fact, these first three types of replication are all forms of technical replication in that they involve the same samples. Therefore, the variance in these measurements is solely due to variability in the experimental process. The last type of replication, and arguably the most important factor in any experimental design, is biological replication. Biological replication refers to hybridizations using samples obtained from different biological specimens. For example, samples derived from individual mice, or separate tissue culture flasks, which have received the same treatment or represent the same condition. Biological replication takes into account the natural variability within the biological population and is necessary to establish that the observed differences in gene expression are due to the experimental condition under investigation, and are not simply an artifact of the biological variance. Microarray experimental designs must take into account practical considerations such as the availability and cost of both microarrays and samples, however whenever possible it is always advantageous to maximize biological replication(Cui & Churchill, 2002)(Pavlidis *et al.*, 2003).

Microarrays are used to quantify gene expression indirectly by measuring the amount of fluorescence emitted by fluorophores attached to transcripts (RNA or cDNA) that have hybridized to specific DNA sequences (probes) that have been spotted at precise locations on the array. Depending on the microarray printing process the probe spot sizes range from 100 to 200 μM and they are spaced 130 and 500 μM center-to-center. Microarrays are scanned at resolutions ranging from 3 to 25 μM using a photomultiplier (PMT)(Packard Bioscience, 2005)(Molecular Devices, 2005) or charge coupled device (CCD)(Applied Precision, 2005) based fluorescent imaging system incorporating lasers and/or filters sets appropriate for each fluorophore. One image is generated for each fluorophore with each pixel representing the magnitude of the fluorescence (reported as 16 bit number) of the area defined by the resolution (usually 10 μM) at a precise location on the slide. Data is extracted from the images through image processing or “spot finding” software that registers the two images, aligns a grid of circles corresponding to the known pattern of probes, and reports aggregate data for the pixels within and surrounding each probe. In this respect, acquiring the final data for a microarray hybridization is actually an image processing problem, for which numerous solutions have been developed(Yang *et al.*, 2001b)(Rahnenfuhrer, 2005)(Qin *et al.*, 2005).

Extracting data from hybridization images is just the first step in performing microarray data analysis. The extracted raw data must then be pre-processed to remove biases in the data before a final analysis can be performed to determine differences in gene expression. These preprocessing steps include background subtraction,

transformation, and normalization and are discussed in subsequent sections. All preprocessing steps serve to reduce experimental variability, and produce one final measure of fluorescence for each probe for each sample used in each hybridization. Initial techniques for analyzing gene expression simply selected differentially expressed genes based on the ratio of the final fluorescence measurements for the two samples for each probe(DeRisi *et al.*, 1996). This technique, however, is unreliable due to the large number of measurements acquired from each hybridization and the substantial variability of any single hybridization(Lee *et al.*, 2000). In addition, simple ratio-based analyses do not provide any measure of statistical significance and cannot accommodate more complex experimental designs such as the loop design(Kerr & Churchill, 2001b).

In effort to provide better estimates of differences in gene expression between samples, and in order to assign a statistical significance to those differences, researchers have developed mathematical models and implemented a variety of statistical methods for microarray data analysis. Kerr et al.(Kerr *et al.*, 2000) first described the use of analysis of variance (ANOVA) in combination with optimal experimental designs incorporating replicate measures, for microarrays. This technique uses a basic additive linear model to account for known sources of variability, thereby improving estimates of differences in gene expression and providing a measure of confidence for those estimates. Wolfinger et al. and Li et al. incorporate mixed models into their ANOVA based analysis(Wolfinger *et al.*, 2001)(Li *et al.*, 2004) and Draghici et al. employs an ANOVA approach using replicate spots to estimate an empirical distribution of the noise(Draghici *et al.*, 2003).

Baldi and Long developed their approach from a fully Bayesian framework based on a t -test with regularized variance estimates and adapted degrees of freedom (Baldi & Long, 2001). Lonnstedt and Speed employ an empirical Bayes approach to define a statistic B by combining gene means and standard deviations with estimates of parameters of a prior distribution based on all genes in a replicated set of experiments (Lonnstedt & Speed, 2002). Wang and Ethier have proposed a generalized likelihood ratio test (Wang & Ethier, 2004) based on the two-component model proposed by Rocke and Durbin. Under this model, the measured fluorescence is a linear combination of normal random variable, which dominates at low intensities, and a lognormal random variable, which dominates at high intensities (Rocke & Durbin, 2001). Tusher et al. developed a non-parametric approach called *Significance Analysis of Microarrays* (SAM), which also uses a modified version of the t -test called the relative difference d in which a small positive constant has been added to the gene specific variance. Genes are ranked by their d value and the significance of genes above a user-defined cutoff is determined based on the number of genes falsely identified as differentially expressed during semi-random control permutations of the datasets (Tusher et al., 2001).

Efron and Tibshirani implement a Wilcoxon's statistic to rank differentially expressed genes and then use an empirical Bayes model approach to estimate the associated distribution (Efron & Tibshirani, 2002). Pan et al. propose a nonparametric statistic, called the mixture model method (MMM), in which the distributions of a t -type test statistic are estimated using finite normal mixture models (Pan et al., 2002). Delmar

et al. and Wenqing He have both recently proposed non-normal mixture models using maximum likelihood to make inferences (Delmar *et al.*, 2005; He, 2004). Delmar et al. utilizes variance to group genes and models the sum of square residuals according to a Gamma distribution (Delmar *et al.*, 2005). Wenqing He uses a spline function (He, 2004) to model the distribution of the Z statistic (Zhao & Pan, 2003) proposed by Zhao and Pan. Smyth et al. proposes a method by which replicate spots on an array are used to improve estimates of genewise variances, thus improving inference methods designed to detect differentially expressed genes (Smyth *et al.*, 2005). Mukherjee et al. and Cole et al. both take a different approach, employing machine-learning algorithms to identify differentially expressed genes (Cole *et al.*, 2003) (Mukherjee *et al.*, 2005).

In an effort to apply many of these powerful statistical techniques to our microarray datasets, an analysis platform was developed with supporting software named CARMA (Computational Analysis of **R**eplicate **M**easures for **A**rrays). In addition to performing ANOVA on microarray datasets that incorporate replication, CARMA also performs all of the necessary steps for calculating differential expression in a microarray data set including importing, transforming, and normalizing the raw data files. The analysis is designed to be easy to apply, require only a basic understanding of the underlying principles, and provides output in both an easy to interpret graphical format and a delimited text file. Each step in the process was chosen for its broad applicability to microarray data, with user-defined parameters tailoring the analysis to each experiment. To demonstrate the utility of our approach, an example analysis of a

microarray experiment designed to characterize gene expression in an aquaporin-1 knockout mouse model is presented.

2.2 Implementation

CARMA was implemented using the *R programming language and environment* (Bates, 2005) because of its “wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc) and graphical techniques”. In addition R is available at no cost and runs on a variety of computing platforms. Under CARMA, analysis begins with data files and user-defined parameters being read from delimited files. Normalization between the two channels of each array and between both channels of all arrays is achieved using the *loess* function, which was chosen because of its ability to perform simultaneous location and intensity dependent lowess normalization. ANOVA is implemented using the *aoV* function utilizing partitioned error for replicates, or the *lme* function for more complicated models, including mixed models. Graphical output and delimited text files are generated to present the results of the normalization, analysis of variance, and the ANOVA contrasts. A dataflow diagram for CARMA is presented in Appendix 1.

In addition software was developed to generate simulated microarray data. This software was developed using Microsoft Visual Basic due to its integrated graphical user interface design functionality. This software was designed to generate simulated microarray data that includes known sources of variability. User defined parameters control both the configuration of the array (grid layout and spacing, spot layout and spacing, and replicate count) as well as the hybridization scheme (number of arrays, dyes,

genes, and experimental conditions). There is also a parameter to specify a percentage of missing data to simulate an incomplete dataset (Figure 2.4).

Form1

Array Layout

	Count	Spacing		Count	Spacing
Set Rows	2	4500	Rows	4	200
Set Cols	2	4500	Cols	4	200

Linear Model Effects

	Count	Effect	Standard Dev
Array	12	0	.4
Dye	2	0	.4
Gene	32	10	1
Variety	6	0	.4
Array x Dye		0	.4
Gene x Array		0	.4
Gene x Dye		0	.3
Gene x Variety		0	.3
Replicate	2	0	.1
Error		0	.2
% Missing Data	.25		

Generate Data

Figure 2.4 Input form for generating simulated microarray for use with ANOVA
 User input form for generating simulated microarray data incorporating known sources of variability. The layout of the virtual microarray is defined using the Set Rows, Set Columns, Rows, and Columns fields. These fields define both the count and spacing of the spots in the array. The hybridization scheme, as well as the magnitude and standard deviation of each of the parameters affecting the simulated intensities are specified using the remaining fields. These simulated \log_2 based intensities are generated based on the assumption of a normal distribution for each parameter.

2.3 Results and Discussion

Example dataset

The microarray dataset used for this manuscript contains measurements of gene expression in kidneys from adult aquaporin-1 knockout and wild type mice (McReynolds *et al.*, 2005) (GEO Accession GSE2402). In brief, RNA from kidney medullae of three aquaporin-1 knockout mice and three wild type mice was reverse-transcribed incorporating an amino modified dUTP, labeled using Alexa Fluor 546 and 647 ester dyes, and hybridized to a custom microarray containing the NIA Mouse 15K cDNA clone set (Tanaka *et al.*, 2000) with each clone printed in duplicate. Slides were scanned using an Applied Precision arrayWoRx Biochip Reader and image analysis was accomplished using softWoRx Tracker software. Because this experiment is balanced for both mouse (each mouse sample is hybridized twice with each dye) and genotype (3 mice of each genotype were used), it was possible to perform two different analyses; one to determine differences in gene expression between mice and the other to determine differences in gene expression between the aquaporin-1 knockout and wild type groups. The results of the comparison between the aquaporin-1 knockout and wild type groups have been reported previously (McReynolds *et al.*, 2005), therefore this paper presents the comparison of gene expression between individual mice in order to demonstrate the functionality of CARMA.

Analysis of variance (ANOVA) and linear model

In the simplest sense, microarray data is the measured intensities, at defined wavelengths, of elements (or “spots”), which have been arrayed on a glass slide. Included in these measurements are numerous sources of variability. Given that a two-channel microarray experiment consists of multiple samples labeled with two dyes hybridized to multiple arrays containing multiple spots (that represent genes), there are four main sources of variability termed: Array, Dye, Gene, and Variety(Kerr *et al.*, 2000). The Array term refers to the variability in the measured intensities associated with a specific hybridization, due to either variability between slides, variability between hybridizations (e.g. differences in the amount of cDNA used in each hybridization), or both. The Dye term refers to variability caused by differences in fluorochrome chemistry, coefficient of extinction, incorporation efficiency, photobleachability, scanner sensitivity, etc. The Gene term refers to each element (or replicate elements) on the microarray. Since each element will hybridize to its complementary sequence in each sample, the intensity of each spot will depend on its nucleotide sequence, which is considered unique. The Variety term refers to the distinguishing feature of interest (such as time, treatment, or dosage) in the experimental samples. The goal of most microarray experiments is to determine the effect of this Variety term on the measured intensity for each element. In other words, how does dosage (or time, treatment, genotype, etc.) affect the expression of each gene?

Based on the four main effects (Array, Dye, Gene, Variety) and allowing for all interactions between those effects, there are 16 possible terms that could be included in the model(Kerr & Churchill, 2001b), however many of the interaction effects do not

make practical sense. For example, the Variety term is completely encompassed by the Array x Dye term, and the Array x Dye x Gene term would obviate the Variety term. In addition, performing a log-based transformation (see later sections) on the microarray dataset before applying the linear model allows the use of an additive linear model. Equation 1.1 describes the collection of mathematical equations that can be used to calculate values for each of the known factors that contribute to the transformed measured intensities:

$$i_{ijkl} = \mu + A_i + D_j + G_k + (AD)_{ij} + (GA)_{ki} + (GD)_{kj} + (GV)_{kl} + \varepsilon_{ijkl} \quad (1.1)$$

Equation 1.1 defines a relatively complete model for the sources of variability in a microarray experiment; however there are practical limitations to its implementation. Utilizing a least squares approach to solve equation 1.1, even for reasonably small microarray datasets (e.g. 4 hybridization of a 10,000 element array), requires gigabytes of memory, precluding the use of a personal computer. In addition, it assumes equal variance between all genes. Splitting equation 1.1 into two equations, one containing all Gene independent terms, and another equation containing all gene-dependent terms, yields equations 1.2 and 1.3 respectively:

$$i_{ij} = \mu + A_i + D_j + AD_{ij} + \varepsilon_{ij} \quad (1.2)$$

$$i_{ijkl} = G_k + GA_{ki} + GD_{kj} + GV_{kl} + \varepsilon_{ijkl} \quad (1.3)$$

Applying these two models sequentially to a dataset significantly reduces the memory requirements and time required for computation, and allows for gene specific variances. Comparisons between analyses employing single equations similar to equation 1.1 and two equations similar to 1.2 and 1.3 have demonstrated that while the

calculated significance for each gene between the two procedures is often different (as a result of allowing gene specific variances), the list of genes identified as differentially expressed is largely the same (Wolfinger *et al.*, 2001). When using these two models, equation 1.2 effectively serves to perform a linear global normalization between the two channels for each array and between arrays. Instead, because only the Gene x Variety effect is usually of interest, and in order to allow for non-linear global normalization, CARMA performs a lowess normalization between the two channels of each array and between both channels of all arrays, followed by a gene-by-gene ANOVA implementing equation 1.3 alone. Limiting equation 1.3 to the subset of data for each gene yields equation 1.4:

$$i_{ijkl} = \mathcal{G}_k + \mathcal{G}A_{ki} + \mathcal{G}D_{kj} + \mathcal{G}V_{kl} + \varepsilon_{ijkl} \implies i_{ijl} = \mu + A_i + D_j + V_l + \varepsilon_{ijl} \quad (1.4)$$

Implementing this form of the equation reduces computation times to a few hours on a basic personal computer (500 MHz Pentium III with 512 MB of memory) for even relatively large data sets (e.g. an experiment involving 32,000 elements and 12 hybridizations).

Background subtraction and transformation

Most microarray scanners generate 16 bit numbers, resulting in measurements ranging from 0 to 65,535 for each pixel in the generated images. Microarray image processing software is then used to analyze each image, generating a multitude of measures for each spot, which are usually further processed to produce one measure of intensity for each channel for each spot. The most controversial part of this process is

whether to subtract some measure of background (defined as the pixels around the areas delineated as spots) from each spot (defined as the pixels within each area delineated as a spot). Background subtraction is often used to reduce the negative impact of local image artifacts, and provide better estimates of gene expression by subtracting fluorescent signal that is unrelated to the fluorescently labeled hybridized target cDNA. Nevertheless, some researchers have concluded that background subtraction increases variance and degrades the performance of subsequent analyses (Qin & Kerr, 2004). In most cases, however, this increase in variance is not simply due to background subtraction, but the combination of background subtraction and a log-based transformation of the data. In effect, subtracting the background intensities from the spot intensities reduces the values for low intensity measurements to the point that error is no longer multiplicative, but additive, causing a log-based transformation to inflate the variance of these small values. In other words, the measurement error associated with small values is not proportional to the true value, but is a combination of some random error added to the true value. For instance, a log base 2 transformation assigns the same significance to the difference between 2 and 32 ($\log_2(32) - \log_2(2) = 4$) as the difference between 2000 and 32000 ($\log_2(32000) - \log_2(2000) = 4$), even though a difference of 32 is well within the noise of current microarray scanning technology and of no real significance.

The logarithm base 2 of the expression ratio is the most widely used transformation for microarray data due to its continuous range of values and similar treatment of up- and down-regulation (Quackenbush, 2002). It has the added effects of improving linearity and variance homogeneity, and converting multiplicative error into

additive error, thus allowing the application of a linear additive model(Kerr *et al.*, 2000). While log transformations work well for moderate to high signal intensities, they have the undesirable effect of magnifying small differences in intensity at low signal intensities(Kerr *et al.*, 2002). In addition, log transformations cannot be applied to negative numbers, which can result from background subtraction, and they transform numbers between 0 and 1 into negative values. To address these problems some researchers have proposed discarding data below a threshold(Yang *et al.*, 2001a), while others have proposed variance-stabilizing transformations(Huber *et al.*, 2002;Cui *et al.*, 2003;Rocke & Durbin, 2003). CARMA implements a version of the linlog transformation(Cui *et al.*, 2003) that has been adapted to better cope with large negative numbers (where the local background signal intensity is significantly higher than the spot signal intensity) as follows.

$$Z_{ik} = \begin{cases} \log_2(-1/Y_{ik}) + 2 * \log_2(d_i) - 2 / \ln(2) & Y_{ik} \leq -d_i \\ \log_2(d_i) + Y_{ik} / (d_i * \ln(2)) - 1 / \ln(2) & -d_i < Y_{ik} < d_i \\ \log_2(Y_{ik}) & Y_{ik} \geq d_i \end{cases}$$

where Z_{ik} represents the transformed intensities, Y_{ik} represents the untransformed intensities, and d_i represents the threshold between the log and linear portions of the transformation. The subscripts i and k denote the array and element (spot) for each intensity, respectively. This modified linlog transformation is symmetrical around 0 (raw signal) and both it and its first derivative are continuous. It is similar to a \log_2 transformation for both large positive and large negative intensities, and a linear transformation at low intensities (positive or negative). Figure 2.5 presents the results of both a \log_2 and linlog transformation for raw intensities between -1024 and 1024.

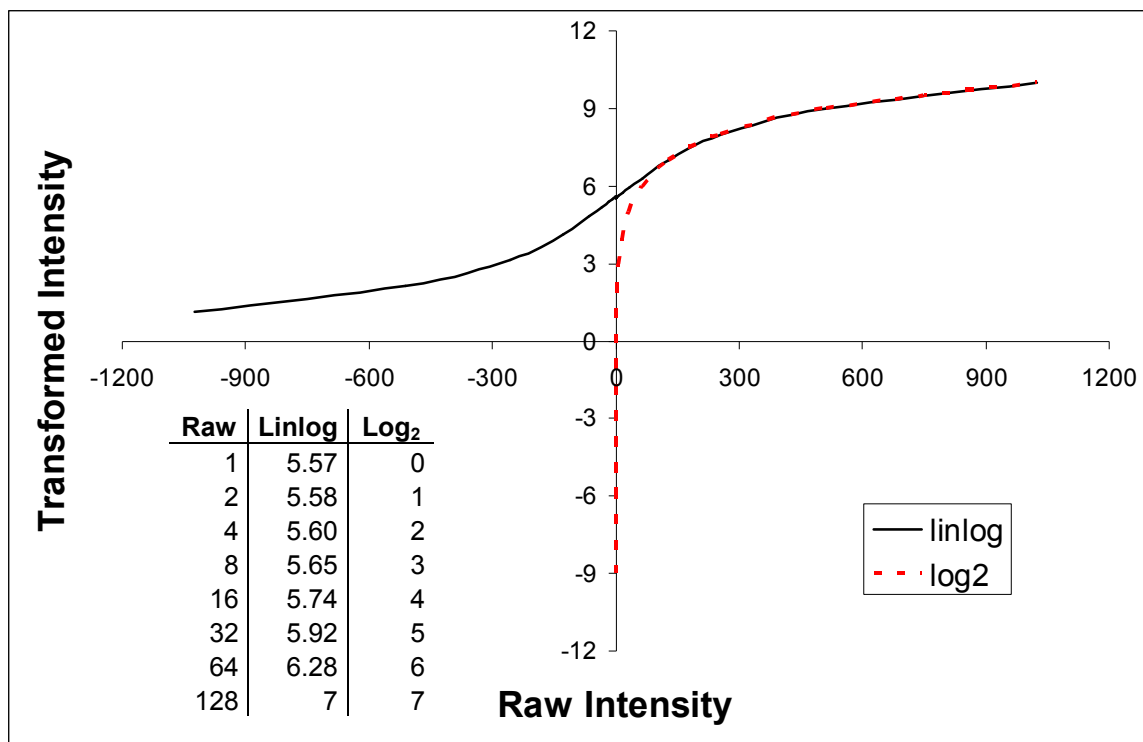


Figure 2.5 Comparison of \log_2 and linlog transformations

Plot of both a \log_2 (red dashed line) and linlog (black solid line) transformation for raw intensities between -1024 and 1024. Note the difference in transformed values between the two transformations for raw intensities less than the linlog threshold, which was set to 128 for this figure. Not only is the \log_2 transformation unable to transform negative values, but it assigns large negative values to raw intensities between 0 and 1. The \log_2 transformation also overemphasizes the significance of small numbers (as indicated by the height of the gray line), which are below the level at which microarray scanners can confidently measure fluorescent signals. The linlog transformation also has the advantages of both it and its first derivative being continuous and it is symmetrical around 0 (raw intensity). The table in the lower left quadrant of the graph highlights the difference between the linlog and \log_2 transformations for intensities below the linlog threshold (above 0).

For the example analysis in this manuscript, the crossover point (d_i) between the linear and log portions of the transformation was calculated based on the median of one standard deviation of the local background of all spots. This method of calculating the cutoff for the linlog function was based on the observation that the variability in the measured intensities at background levels provides a good indication of the minimum intensity that the scanner can accurately measure (the point at which error becomes multiplicative). This approach has worked well in practice, and has the advantage of not assuming any distribution for the data. Recently CARMA has been enhanced to include the capability of calculating the linlog crossover point based on minimizing the absolute deviation of the inner quartile range (IQR) for each bin from the median IQR, for 20 bins spanning the range of intensities for each hybridization.

Utilizing the aquaporin-1 dataset, a comparison of the effect of local background versus no background subtraction, and the consequences of applying either a \log_2 or linlog transformation, on the ability to detect differentially expressed genes was conducted. Because this dataset consists of four replicate measurements for 6 mice, one success criterion is minimizing the within mouse variance while maximizing the between mouse variance (Ding & Wilkins, 2004), which equates to maximizing the ANOVA F value for the Variety (mouse) term, in equation 1.4. In the case of \log_2 background subtracted data, values less than 1 were set to 1, and in the case of the linlog transformation, the crossover point between the linear and logarithmic segments was set to the median of one standard deviation of the background for all spots. Figure 2.6 illustrates the effect of each background subtraction/transformation method on one of the

hybridizations in the dataset. The most obvious change due to background subtraction is the expansion of the range of the data resulting from the removal of the relatively large signal floor associated with the longer exposure times and CCD-based image capture employed in the Applied Precision arrayWoRx Biochip Reader. This large signal floor also negates any difference between the \log_2 and linlog transformation on the non-background subtracted data because error remains multiplicative across the entire range of values. The difference in transformation applied to the background subtracted data however, is obvious as indicated by the larger spread of the lower range of the \log_2 transformed data as compared to the linlog transformed data. Table 3.1 presents a summary quantitative comparison of the effect of each transformation on the ratio of the between group mean squares over the within group mean squares after location and intensity dependent lowess normalization and gene-by-gene analysis of variance. On a row-by-row basis after ranking, the ANOVA F values for the linlog transformed local background subtracted data were larger than any of the other combinations of background subtraction and transformation, for every element in the dataset. Also, applying a step-up p -value procedure (Benjamini & Hochberg, 1995) to the ANOVA p -value controlling at a 5% false discovery rate resulted in 8 genes being identified as differentially expressed for both of the non-background subtracted datasets, and 22 and 129 genes being identified as differentially expressed for the \log_2 transformed and linlog transformed background subtracted data, respectively.

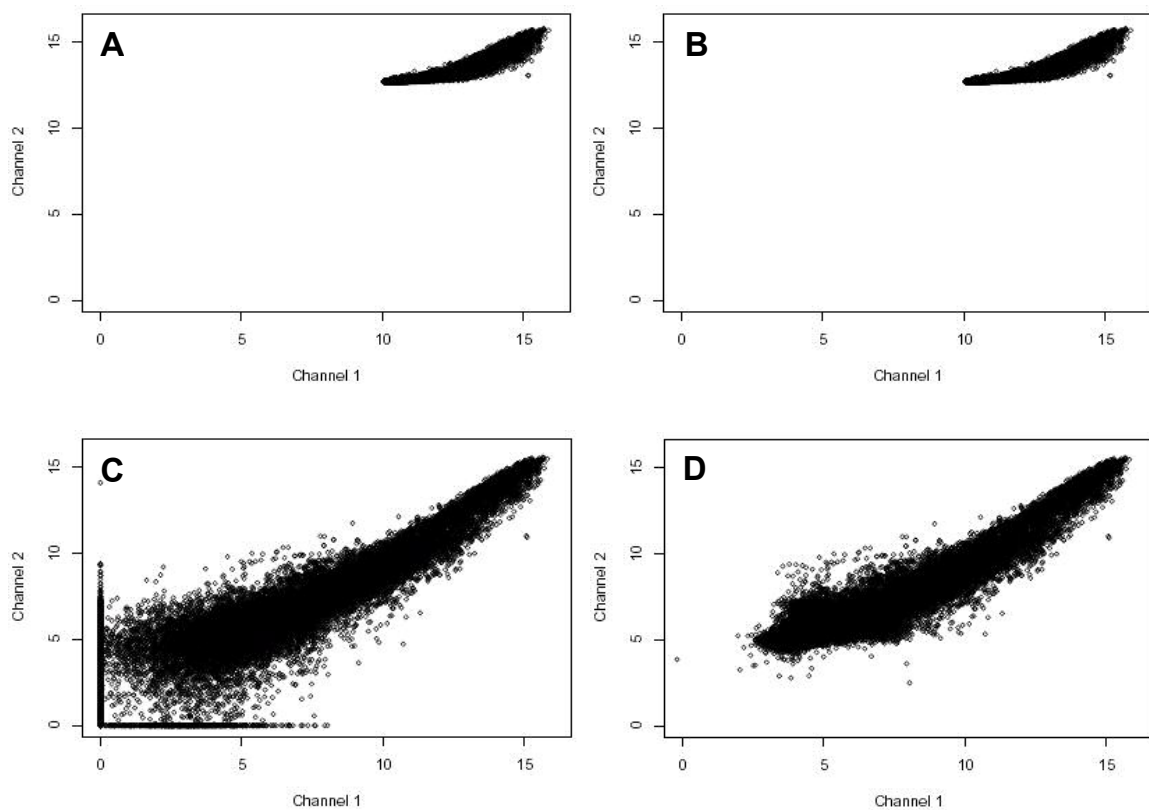


Figure 2.6 Comparison of pre-normalized and post-normalized data

Plot of the pre-normalized data, channel one (green) vs. channel two (red), for one hybridization in the aquaporin-1 dataset after applying **A** log₂ transformation without background subtraction or **B** linlog transformation without background subtraction or **C** log₂ transformation after background subtraction or **D** linlog transformation after background subtraction.

Normalization

Differences in fluorochrome characteristics, scanner wavelength sensitivities and settings, dye incorporation, and other non-biological effects all contribute to differences in the intensities between the two channels of any hybridization. And while global normalization techniques apply the same adjustment to every spot for a hybridization(Quackenbush, 2002), it is often necessary to correct for intensity and location specific effects(Yang *et al.*, 2002b;Yang *et al.*, 2002a;Cui *et al.*, 2003;Smyth & Speed, 2003;Wilson *et al.*, 2003). CARMA implements a locally weighted regression (lowess)(Cleveland, 1979;Yang *et al.*, 2002b) transformation that can adjust for either intensity or location (or both) dependent effects. In addition, CARMA normalizes between both channels of all arrays, serving to minimize the Array term in the linear model and aiding in the visualization of the normalized data.

Figure 2.7 displays the data for one of the hybridizations in the aquaporin-1 dataset, both before and after normalization. As is common with two dye hybridizations, pre-normalized data shows curvature of the data at lower intensities (Figure 2.7a). Following lowess normalization for both intensity and position on the array, the curvature (intensity bias) of the plotted data is removed and the distribution of the data is narrowed (spatial bias) as compared to the pre-normalized data (Figure 2.7d). Our observations indicate that the location dependent effect is due to spatial hybridization variability and spatial scanning biases often caused by photo bleaching. In practice, the location of an element on the array can play a significant role in normalization, particularly with epifluorescence-based scanners (unpublished observation), and can vary by as much as a

factor of 2 from one end of the slide to the other (Figure 2.7e). The extent to which location affects the intensity measurements is specific to each hybridization and scanning process.

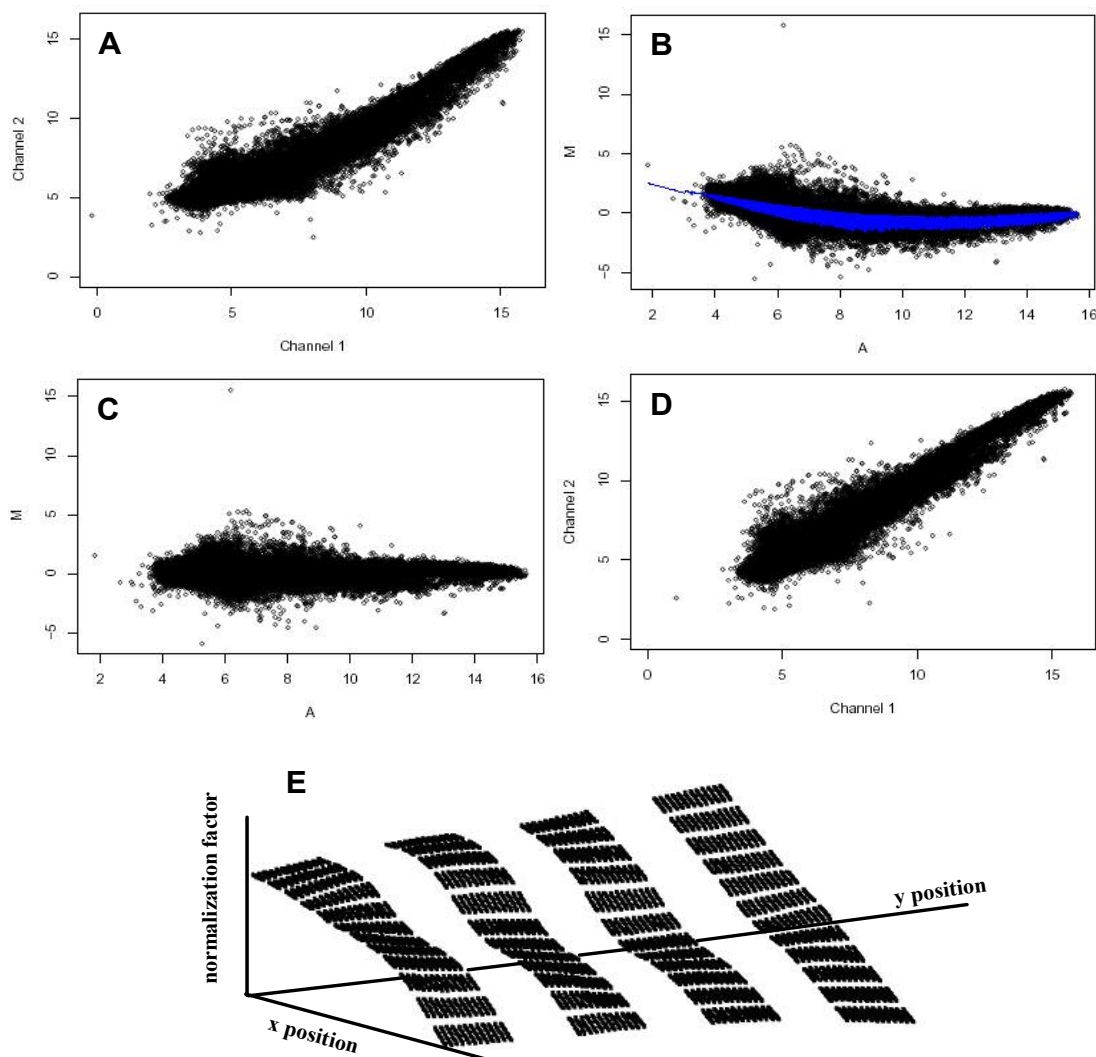


Figure 2.7 Normalization of a two-channel hybridization

A The linlog transformed data before normalization plotted as channel 1 (Alexa 546 dye) versus channel 2 (Alexa 647 dye) intensities. **B,C** Ratio-Intensity plot before (**B**) and after (**C**) spatial and intensity lowess normalization. “M” refers to the log ratio of the two channels and “A” refers to the geometric mean of the spot intensity for both channels. The blue line in **B** is the spatial/intensity lowess normalization fitted curve, with the curvature of the line representing the intensity-dependent fit and the width of the line representing the spatial component of the lowess normalization. **D** The final normalized data used for the ANOVA. **E** A plot of the correction factors, by array position, used to normalize intensities due to spatial effects only. Shown are the amount adjusted for each of the 650 elements in the 12 X 4 subarrays (in the X and Y directions, respectively) of the mouse cDNA microarray used in the study. A location dependent effect can be observed as a general increase over the length of the slide (bottom to top) and a dip near the center of the slide.

Variance shrinking

Microarray datasets usually contain thousands of elements, each representing one gene, but only a few measurements for each element. Whereas performing a global ANOVA on the entire dataset assumes equal variance within the data for each gene, it is generally accepted that independent analysis of the subset of data for each gene does not utilize sufficient data to determine an adequate representation of the variance associated with each gene. On the other hand, the major consequence of performing a gene-by-gene ANOVA is the overestimation of the significance of the calculated differences in expression for genes with abnormally small variances, and the underestimation of the significance of the calculated differences in expression for genes with abnormally large variances. Researchers have addressed these issues by including information from all genes on the array when assessing the significance of differences in expression for each gene (Baldi & Long, 2001; Lonnstedt & Speed, 2002; Kendzioriski *et al.*, 2003; Wright & Simon, 2003; Smyth, 2004). CARMA calculates four variances and associated p-values for each gene (Cui *et al.*, 2005): gene specific variance, pooled variance (average for all genes), half of the gene and half of the pooled variance, and an estimator based on the James-Stein-Lindley shrinkage concept (Lindley, 1962) that uses a formula to calculate the variance based on both the gene and pooled variances.

Data filtering and outlier detection

Given that the samples hybridized to most large microarrays will contain transcripts for only a subset of the genes represented on the microarray, removing spots

that exhibit low intensities for all samples can both reduce the number of genes incorrectly identified as differentially expressed between samples and decrease the computing time required for subsequent analyses. As implemented in CARMA, the ANOVA is usually (based on typical user settings) performed on only those genes that have background subtracted intensities greater than 1 or 2 background standard deviations for at least 51% of the measurements for at least one sample in the hybridization scheme. Furthermore, in the case of microarrays with replicate elements, usually at least 51% of the replicates must have background subtracted intensities greater than 1 or 2 background standard deviations in order for any of the replicates to be included in the ANOVA. In other words, the ANOVA is only performed on genes that are consistently expressed at measurable levels in at least one of the samples.

CARMA also has the capability of removing anomalous measurements through outlier detection. Dust, impurities, surface inhomogeneities, fluorochrome-specific effects, local hybridization effects, technician effects, etc. all contribute to non-systematic variability in the measured intensities. Anomalous measurements can be identified by their incongruity with other measurements within a hybridization and through inconsistencies between replicated measures (Yang *et al.*, 2002a; Quackenbush, 2002). Following the ANOVA of the normalized data for a gene, CARMA applies the following formula to identify outliers:

$$r_{ii} > |quantile((OP / 2) / n, df - 1)|$$

Where r_{ii} = studentized residual of the i th element, OP = user defined outlier probability, n = the number of measurements, and df = degrees of freedom. ANOVA is performed

recursively, removing the most extreme outlier after each step, until all outliers (which meet the criteria above) have been removed. The results of both the original ANOVA (using all data points), and the last ANOVA, performed on the dataset with all outliers removed, are shown in the graphical output (Figure 2.8A).

Missing data

Most microarray hybridizations will have regions that are obviously problematic. While it may not be worthwhile to flag small abnormalities such as fine dust particles, large abnormalities should be flagged for exclusion from analysis. Missing data, whether flagged manually, filtered out, removed through outlier detection, or unavailable because of a failed hybridization may result in some experimental imbalance. This imbalance, however, is less problematic than including erroneous data. Therefore, CARMA utilizes functions that can accommodate missing data, to apply the linear model and perform the ANOVA.

Output and display

Any analysis is only as good as its ability to provide accurate relevant information to the researcher. In addition to generating tab delimited output files, CARMA creates an Adobe Portable Document Format (pdf) file containing easy to interpret graphical output (Figure 2.8) including an estimate, and its standard error, for each level (possible value) of the Variety effect, as well as plots of the normalized data and other statistical quality control information. This graphical output allows the researcher to refine and prioritize

the list of differentially expressed genes by helping to identify cases of non-normality, outliers, unexpected patterns, etc. Visualizing the normalized data and the results of the ANOVA also helps to identify and correct cases of mislabeled samples and misaligned grids. All of the numbers used and generated by the ANOVA, and a file containing the contrasts between all pair-wise combinations of the levels of the Variety effect are also provided in delimited text files.

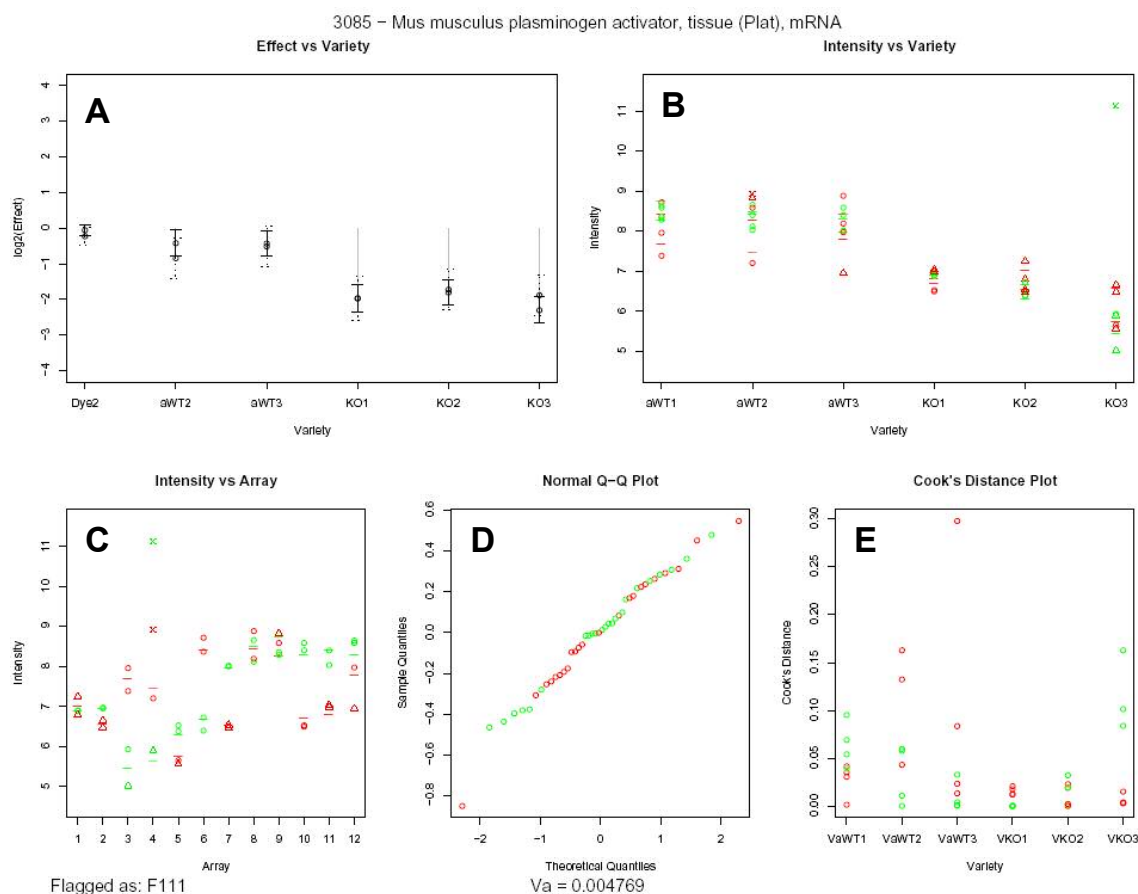


Figure 2.8 Results of ANOVA for the Mus Musculus *Plat* gene

In panels B-E of this figure the color of the plotted data points represents the fluorochrome that was used to label the sample (green = Alexa 546, Red = Alexa 647). **A** Graphical display of the Variety term estimate and standard error for the relative *Plat* gene expression for mice WT2, WT3, KO1, KO2, KO3 referenced to WT1 (the a in front of WT1, WT2 and WT3 is simply a label marker). Solid lines represent the final estimates after removal of outliers, while the dashed lines represent estimates before removal of outliers. The Dye2 (Alexa 647) effect and its standard error are also shown. **B,C** Transformed and normalized intensities plotted by sample (**B**) or hybridization (**C**). Colored circles (confident) and triangles (below user defined confidence threshold) represent the normalized measured intensities for each element (i.e. spot), and dashes represent the calculated intensities from the ANOVA model. An x denotes a point that was identified as an outlier. **D** A normal Q-Q plot for all data providing an indicator of the normality of the residuals. **E** The Cook's distance plot illustrating the influence of each data point on the fit of the model.

2.4 Conclusion

Microarray experiments are intended to determine relative differences in gene expression between various treatments or conditions. As with nearly all experiments, and exacerbated by the number of measures obtained from microarray hybridizations, experimental noise can confound measurements and lead to the incorrect identification of random variations as significant differences in gene expression. We have employed a generalized approach and developed supporting software (CARMA) for performing ANOVA on microarray datasets that is easy to implement, generates readily interpretable graphical results, and accounts for experimental sources of variability. Applying ANOVA to microarray datasets incorporating replicated measures improves estimates of differences in gene expression between samples and provides a statistical basis for determining the significance of these differences. Also, because each sample is involved in multiple hybridizations, it is possible to identify and remove incongruous data points that are caused by dust particles, local background, etc, as well as allow for missing data caused by occasional failed or abnormal hybridizations. CARMA provides a clear quantitative and statistical characterization of each measured element on the microarray that can be used to assess marginally acceptable measures and improve confidence in the interpretation of microarray results. Overall, applying CARMA to microarray datasets incorporating replicated measures effectively reduces the number of gene incorrectly identified as differentially expressed and results in a more robust and reliable analysis.

In some situations it is advantageous to study the effects of more than one Variety on gene expression. For example, in gene knockout studies both genotype and treatment

dependent effects may be important. Also, in cases where subjects are exposed to multiple treatments, accounting for individual-specific effects may expose differences in gene expression due to treatment that might otherwise be obscured. In these cases equation 1.4 can be modified to include terms for each of the varieties of interest. For example, the following model could be used to examine the effects of genotype (V1), drug treatment (V2), and the interaction between genotype and drug treatment (V1V2):

$$i_{ijlm} = \mu + A_i + D_j + V1_l + V2_m + V1V2_{lm} + \varepsilon_{ijlm} \quad (1.5)$$

Theoretically equation 1.5 could be expanded to include any number of varieties, however because of an exponential increase in the number of hybridization that must be performed most researchers limit their experiments to studying a maximum of two varieties.

Implementation of CARMA depends on a few assumptions. First, the lowess normalization assumes that the expression levels for the majority of genes in each sample are the same. While this is usually the case, even under conditions where the expression levels of the majority of genes are affected, lowess normalization produces the desired result of adjusting the dataset such that genes that behave dissimilarly from the majority are more likely to be identified as differentially expressed. Second, when determining the significance of differences in gene expression between samples, it is assumed that the data is normally distributed after the application of the linear model. This assumption is consistent with our experience with a variety of microarray datasets as well as that of other researchers (Wolfinger *et al.*, 2001). In addition, a normal Q-Q plot is generated for each gene to provide an indication of the validity of this assumption on a gene-by-gene

basis. We make this assumption in order to utilize the readily available R packages that perform the ANOVA and their superior computational efficiency over non-parametric approaches. Lastly, in order to apply the ANOVA to each gene individually, it is assumed that after transformation and normalization each gene is independent of the other genes on each array. Theoretically, because each element (gene) is on every array there is not complete independence between genes, however we chose to implement the ANOVA on each gene independently in order to allow genes to have dissimilar variances, and to efficiently implement the ANOVA.

Background subtraction has been criticized for inflating the variance of microarray datasets(Qin & Kerr, 2004), but this increased variability is almost always limited to elements that exhibit low intensity measures. This increased variability is not simply due to the smaller numbers, but rather the methods by which these numbers are measured and transformed. Both photomultiplier tubes (PMT) and charge-coupled devices (CCD) cannot accurately distinguish small differences in intensity. Therefore the assumption that error is multiplicative for these values is inaccurate(Kerr *et al.*, 2002) and thus the application of a basic log-based transformation is inappropriate. The commonly seen flaring of the data at lower intensities, in either a simple green channel vs. red channel or a ratio-intensity plot, is not usually an indication of the inappropriate application of background subtraction, but rather an inappropriate transformation. In addition, not removing the relatively large uniform background associated with some datasets, such as those generated by CCD based scanners, can completely obscure the magnitude of any differences in gene expression. As illustrated by this dataset, different

approaches to background subtraction and transformation can have a significant effect on the identification of differentially expressed genes. In fact, at most only 25 % the top 100 genes were shared between the background subtracted and non-background subtracted results.

Many statisticians are hesitant to remove outliers based solely on statistical criteria. In the case of microarray datasets, however, we know that there are invalid measurements in every dataset; yet it is often impractical to manually flag each instance. CARMA not only automatically detects and removes outliers, but also provides supporting graphs to assist in the final determination of the validity of the measurements that were removed. For example, in Figure 2.8, panels B and C, the green 'x' that indicates one of the excluded outliers is clearly separated from the rest of the measures (the corresponding red measurement is also excluded). Of course the researcher can also investigate each spot on the images from which the measurements were derived to further authenticate measurements flagged as outliers, or turn off outlier detection altogether if so desired. In addition to removing sporadic anomalous measurements, outlier detection has proven invaluable for detecting mislabeled samples and misaligned grids. In these cases, without outlier detection, all or a part of a microarray dataset is often labeled as too variable and of little value. A quick inspection of CARMA's graphical output reveals these problems as a series of genes for which the same hybridization's values were dropped as outliers, providing not only an indication that there is a problem, but also the exact hybridization that is the cause of the problem.

Because of the balanced design of this experiment we were able to assess the affects of inter-individual variability in the aquaporin-1 knockout vs. wildtype microarray dataset. The fact that 3% (129 out of 4361) of the confidently measured genes were identified as differentially expressed between mice of the same genotype highlights the significant amount of variability in gene expression between even genetically similar mice. This finding corroborates the results of an earlier study(Pritchard *et al.*, 2001), underscoring the need for including biological replicates in any study, especially those addressing gene expression and molecular activities.

CARMA was designed to be an integrated, easy to use, analysis platform that researchers can apply to their microarray datasets without preprocessing their data or writing any computer code, with an emphasis on providing results in an easily interpretable format. In particular, we have attempted to identify and address issues such as the relatively high background and scanning related photobleaching that can occur with CCD based imaging systems, and include means to address many of the real-world problems associated with microarray experiments. There are already a number of existing microarray analysis tools that prove useful in analyzing microarray datasets (Smyth, 2004;Gentleman *et al.*, 2004;Parmigiani *et al.*, 2003). CARMA implements many of the same statistical and analytical processes employed in these packages and includes the following additional features:

- Ability to read data files generated by most microarray image processing software
- No need to preprocess or combine raw data files
- Automatic exclusion of genes with low confidence measures for all samples

- Modified linlog transformation that better handles large negative numbers (that can result from background subtraction)
- Automatic computation of linlog crossover point
- Simultaneous intensity and location lowess normalization
- Ability to process incomplete datasets for fixed effect models
- Automatic outlier detection and removal
- Detailed graphical output for each gene
- Ability to detect and identify misaligned grids or mislabeled samples

The relative expression values generated by CARMA (i.e. the Variety value for each gene relative to a reference sample), can be further processed by commercial and freeware software packages designed to organize, cluster and display microarray data. In this regard, the relative expression values are analogous to the more traditional ratio-metric measures in that both provide a relative difference in expression between samples. Of course more advanced users can modify or extract portions of CARMA to integrate into their own analyses if so desired. Future development of CARMA may include developing a graphical user interface, improving the implementation of mixed models, adding new methods of normalization, implementing bootstrapping to calculate significance, and adapting CARMA to function as a Bioconductor package (Gentleman *et al.*, 2004).

3 AGGLOMERATIVE HIERARCHICAL CLUSTERING

3.1 Introduction

Identifying differentially expressed genes is the first crucial step in the analysis of any microarray dataset, however the steps that follow depend on the objective of the experiment. Early microarray experiments were often undertaken to simply identify differences in gene expression between two samples (Schena *et al.*, 1996). The most common experimental design involved merely hybridizing the labeled cDNA from a control sample against labeled cDNA from a treatment sample. For example, one sample might be from a flask of cells that had been treated with a drug and the other sample from an untreated flask. These early experiments often employed custom cDNA arrays that contained unsequenced clones from a cDNA library. Genes would be ranked based on the magnitude of the log ratio of the measured intensities, and the genes that exhibited the biggest differences in expression would be sequenced for identification. Northern blots or semi-quantitative RT-PCR would then be used to validate differences in gene expression. In essence these early experiments were gene discovery experiments designed to find novel genes associated with a variety of difference conditions (e.g. environment, treatment, genetic background, time, tissue, etc.).

As microarray technology progressed and microarray experiments grew larger so did the types of analysis that could be performed. Experiments involving multiple samples, such as a time course or dosage response, provide sufficient measurements to group genes by their expression profiles. This technique, known as “clustering”,

organizes genes into groups (or clusters) based on the similarity of their expression measurements for each sample (Eisen *et al.*, 1998). Clustering provides an intuitive interpretation of microarray data based on the understanding that genes that exhibit similar expression profiles are more likely to be coordinately regulated and often involved in the same processes.

Clustering has also been applied to microarray datasets in an effort to classify samples, instead of genes, based on their gene expression profiles. In particular, researchers are particularly interested in classifying tumor samples. Current cancer diagnosis is based on a combination of clinical, histopathological, and molecular parameters, however atypical clinical presentation or morphological irregularities often complicate the classification of malignancies. In addition, patients with the same diagnosis often display dramatically different responses to equivalent treatments. While the exact cause of this variation in response is unknown, it is well established that molecular heterogeneity caused by chromosomal translocations, deletion of tumor suppressor genes, and changes in chromosomal number exist within individual cancer diagnostic categories. In one of the first studies of its kind, Alizadeh *et al.* utilized a cDNA microarray to perform a molecular classification of B-cell lymphomas, which not only distinguished between the known subtypes of malignancies, but also identified a previously undetected subtype that was indicative of better overall survival (Alizadeh *et al.*, 2000). Ramaswamy *et al.* went one step further by demonstrating the ability to distinguish between 14 common tumor types originating from a variety of tissues (Ramaswamy *et al.*, 2001). Microarray experiments have also been used to predict

clinical outcome based on gene expression profiles (Modlich *et al.*, 2005) (van't Veer *et al.*, 2002).

Clustering and classification algorithms can be divided into two main categories: supervised and unsupervised (Rahnenfuhrer, 2005). Unsupervised clustering algorithms do not require any input regarding true cluster membership. In other words there is no a priori knowledge of actual or suspected group membership. These techniques are mainly used to group genes based on expression profile. Supervised clustering techniques involve utilizing known or suspected group membership to “train” a clustering algorithm. The trained algorithm is then used to confirm group membership or group data for which no cluster information is known. These techniques are usually used to classify genes or samples. The first and still widely used method of clustering microarray data is an unsupervised technique called agglomerative hierarchical clustering (Eisen *et al.*, 1998). This process recursively groups genes into clusters beginning with all genes being unclustered and ending with all genes belonging to one cluster. A second closely related hierarchical clustering technique which has also been applied to microarray data is divisive clustering. Divisive clustering proceeds in the opposite order of agglomerative clustering with all genes belonging to the same single cluster at the beginning of the process, which is repeatedly divided in two until all genes are singular (Alon *et al.*, 1999). Notterman *et al.* also introduced the concept of two-way clustering in which both the genes and samples are clustered concurrently (Notterman *et al.*, 2001). An example of an agglomerative hierarchical tree is presented in Figure 3.1.

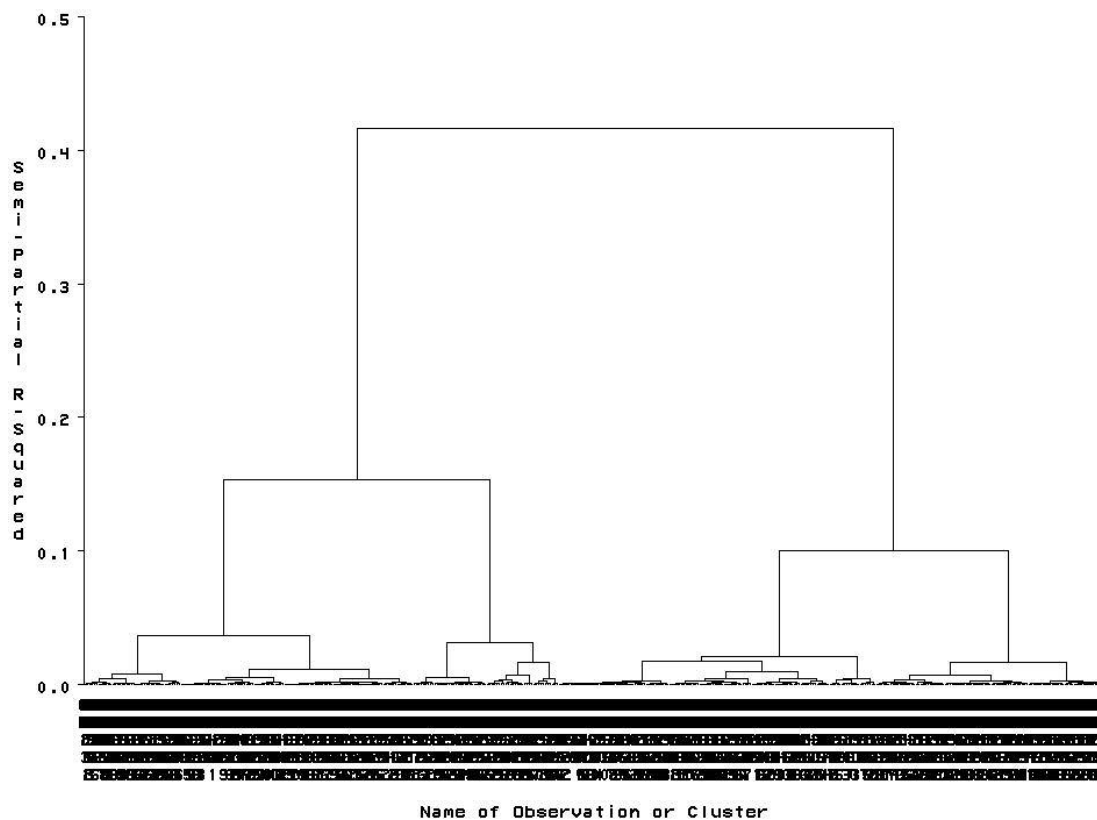


Figure 3.1 Agglomerative hierarchical clustering tree

An example hierarchical tree that displays the distance (y-axis) between each gene and/or cluster (x-axis).

Other techniques for unsupervised clustering of microarray data include self-organizing maps (SOMs)(Toronen *et al.*, 1999) and k-means(Sherlock, 2000) clustering. These clustering techniques are different from hierarchical clustering in that the number of clusters must be specified upfront. SOMs work using a network of nodes that adjust to best fit the pattern of data created by mapping microarray data onto an N-dimensional geometric space, where N represents the number of samples in the experiment. After an initial random seeding of the nodes, each gene is evaluated and the node closest to the gene is adjusted to better match the expression profile of the gene. In addition, adjacent nodes are also adjusted, although to a lesser degree. This process is repeated thousands of times, with subsequent iterations resulting in smaller adjustments, until the network stabilizes (Figure 3.2). At this point the genes are grouped into clusters based on the node to which they are closest.

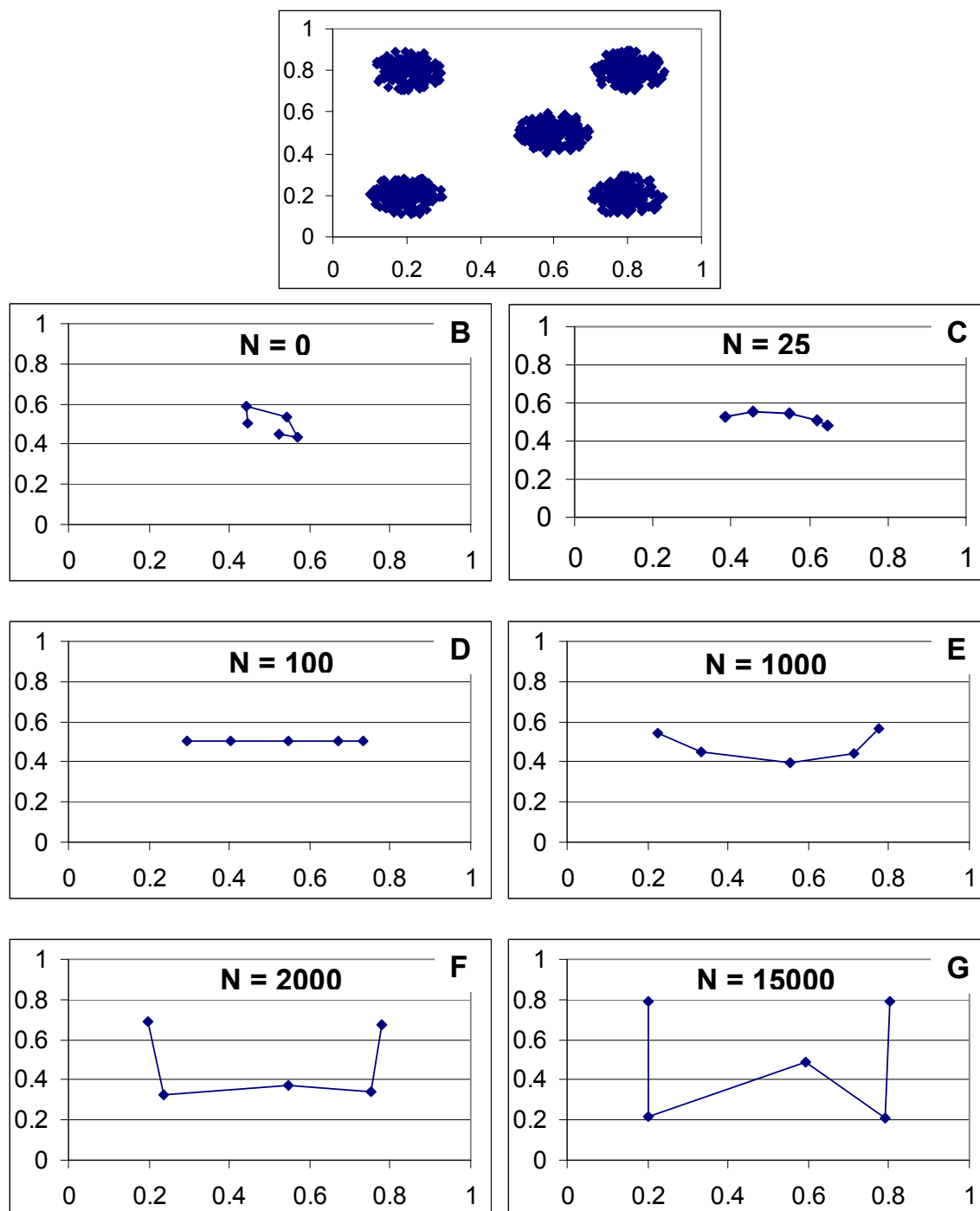


Figure 3.2 One-dimensional self organizing map

Example of a one-dimensional self organizing map containing 5 nodes adapting to a two dimensional input dataset containing 5 well separated clusters (A). The location of the nodes within the two dimensional space is shown after 0 (B), 25 (C), 100 (D), 1000 (E), 2000 (F), and 15,000 iterations (G). After 15,000 iterations the nodes are arranged such that each of the nodes corresponds to a cluster in the input dataset.

K-means clustering is similar to a SOM in that a set of reference vectors (similar to nodes) is initially created through random seeding, however in k-means clustering the reference vectors do not influence each other. After initial seeding each gene is assigned to its closest reference vector, which is then recalculated to include the new gene's expression profile. This process repeats until all genes have stabilized, creating clusters of genes associated with each reference vector. Heyer et al. proposed another unsupervised algorithm in which genes are assigned to clusters whose maximum diameter is specified (Heyer *et al.*, 1999). This technique avoids having to specify the number of clusters up-front and also allows for genes to remain unclustered, however it forces clusters of similar diameter (or size).

Cheng and Church devised a new type of clustering designed to detect groups of genes that exhibit a similar expression profile for a subset of samples (Cheng & Church, 2000). Their proposed algorithm, which allows for overlapping groupings of genes, looks for submatrices within a microarray dataset that exhibit low mean squared residual scores. Tanay et al. combine graph-theoretic and statistical modeling to detect biclusters of genes that jointly respond across a subset of conditions (samples) (Tanay *et al.*, 2002). To model microarray data they use a bipartite graph whose two parts correspond to genes and conditions, which are connected with edges representing significant differential expression. Tanay et al. went on to apply their biclustering technique to a highly diverse collection of genome-wide datasets, thus revealing numerous biological processes using heterogeneous sources and type of data (Tanay *et al.*, 2004). Other approaches to

biclustering include Gibbs sampling(Sheng *et al.*, 2003) and singular value decomposition(Kluger *et al.*, 2003).

Hierarchical clustering, SOMs, and k-means clustering are still the most widely used methods for summarizing the results of microarray experiments because of their intuitive interpretation and relative ease of implementation. Inherent to all of these techniques, however, are difficulties in determining the optimal number of clusters and an inability to assess cluster quality. In an effort to estimate the number of clusters in a dataset, Horimoto and Toh utilized a statistics-based value, known as the variance inflation factor, evaluated at each node in the hierarchical dendrogram(Horimoto & Toh, 2001). Dudoit and Fridlyand took a different approach, using a prediction-based resampling method that combines ideas from discriminant and cluster analysis to calculate the optimal number of clusters(Dudoit & Fridlyand, 2002). Ben-Hur et al. also use a sampling technique to assess the number of clusters in a dataset. In their approach, the optimal number of clusters is chosen based on the transition from stable to unstable clusterings of sub-samples of the dataset(Ben-Hur *et al.*, 2002).

To assess the reliability of cluster analysis Kerr et al. employ bootstrapping(Kerr & Churchill, 2001a). This process involves generating a number of simulated datasets in which the measured intensities are perturbed based on calculated residual errors. These simulated datasets are then clustered, and cluster stability is assessed based on the percentage of the time that each gene groups in each cluster. Levine and Domany and Smolkin and Ghosh take resampling approaches to assess cluster validity in which subsets of data are constructed and then clustered(Levine & Domany, 2001)(Smolkin &

Ghosh, 2003). Ben-Hur et al. also use resampling to assess cluster stability, however in their case it is based on the distribution of pairwise similarities between clustering of subsamples of the data (Ben-Hur *et al.*, 2002).

Principal component analysis (PCA) is another unsupervised technique for reducing the complexity of microarray data, however instead of clustering genes into groups sharing similar expression profiles, PCA projects the data onto a lower dimensional space (Armstrong & van de Wiel, 2004). This process, also referred to as singular value decomposition, reduces the data from a matrix of size gene x array to a matrix of size “eigengene” x “eigenarray” (Alter *et al.*, 2000). This new square matrix, whose size is the minimum of the number of arrays and number of probes (genes) per array, contains unique orthonormal superpositions of the genes. The process of generating eigengenes involves calculating the linear combination of genes that explains the maximum amount of variability in the dataset, and then removing the variability from the dataset. This process is repeated until all eigengenes have been calculated, thus each subsequent eigengene explains less and less of the variability in the original dataset. Eigengenes are then often mapped back to biological processes by examining the contribution of each gene to each eigengene. Because successively less variance is explained by each eigengene, only the first few eigengenes are usually considered significant, and examination of these significant eigengenes can reveal global patterns of expression. PCA has also been used to filter out the noise in microarray experiments by removing the insignificant eigengenes (Dewey & Galas, 2001), and it has also recently been used to identify differentially expressed genes (Wang & Gehan, 2005).

As opposed to unsupervised clustering methods that aim to group genes or samples based solely on patterns present in the dataset, supervised clustering and classification techniques attempt to assign genes or samples into predefined classes. These supervised classification techniques are based on the concept of “training” in which some or all of the samples or genes are used to train a model to distinguish between two or more classes. During this training process the model is recursively adapted to maximize the number of genes/samples classified correctly, and minimize the number incorrectly classified. Once the model has been trained it can then be used to classify unidentified samples. In addition, the trained model can be used to identify genes/samples in the training dataset that may have originally been assigned to the wrong class.

Support vector machines (SVMs) are one of the first supervised algorithms to be applied to microarray data. Under SVMs each vector in the gene expression matrix can be thought of as a point in an m -dimensional space. A binary classifier is then constructed as a hyperplane that maximizes the separation between the class members and non-members. Brown et al. first applied a SVM to microarray data in an attempt to classify genes in budding yeast into function groups based solely on gene expression (Brown *et al.*, 2000). The assigned class for each gene was based on classes from the Munich Information Center for Protein Sequences Yeast Genome Database. A more common use of SVMs is to classify samples, especially tumor samples. Furey et al. demonstrated the use of an SVM to classify ovarian cancer, ovarian, and normal tissue samples and in the process identified one sample as mislabeled and another as an

outlier(Furey *et al.*, 2000). As an extension to the binary SVM, Lee and Lee developed the Multicategory SVM (MSVM) to distinguish between more than two classes, which avoids many of the problems associated with combining multiple binary SVMs(Lee & Lee, 2003).

Another technique useful for confirming class membership, rather than defining class membership, is leave-one-out cross validation. This process involves setting aside one gene/sample, training a classification model based on the remaining genes/samples, and predicting the class of the gene/sample that was left out based on this new model. This process is repeated once for every gene/sample, and the total dataset misclassification error rate is calculated based on the percentage of genes/samples whose predicted class matches their given class(Dettling & Buhlmann, 2002).

As stated earlier, one of the first and still most widely used approaches to extract meaningful information from the large data sets generated by microarray experiments is agglomerative hierarchical clustering(Eisen *et al.*, 1998). Clustering reduces microarray data sets by grouping genes or experiments by expression profile, facilitating data visualization, classification, and further analysis. Agglomerative hierarchical clustering proceeds by first calculating an upper-diagonal distance matrix for every pair of n genes for a series of m measurements. This distance is defined by some measurement of dissimilarity, such as Euclidean distance or angle, between the m -dimensional vector that represent the m measurements for each of the n genes being compared. The smallest value in the distance matrix is selected and the two corresponding genes are joined to form a cluster. The distance matrix is updated with this new cluster, and the two

members of the newly created cluster are flagged as clustered. This process is repeated, with subsequent iterations including these newly created clusters in the distance matrix. Clusters that contain other clusters are created in the same manner as clusters that contain two genes. This process of pair-wise combination of genes and clusters is repeated $n-1$ times until only a single cluster remains.

Differences between agglomerative hierarchical clustering techniques fall into two main categories. The first category is the method by which the distance (or dissimilarity) between two genes is calculated (examples include Euclidean distance, squared Euclidean distance, Manhattan distance, and a metric based on Pearson's correlation coefficient). The second category is the manner in which the distance between clusters is determined. A comparison of nearest neighbor, also referred to as single linkage (McQuitty, 1957), and pairwise average linkage (Sokal & Michener, 1958) illustrates this type of difference. Nearest neighbor defines the distance between two clusters as the distance between the two closest members (i.e. smallest distance), one from each cluster. Pairwise average linkage defines the distance between two clusters as the average distance of each pairwise comparison of the members in each cluster.

In an effort to better understand the effects of employing different distance metrics and clustering methodologies on the clustering of cDNA microarray data we developed a generalized computer algorithm to implement 10 hierarchical clustering methods and 4 distance metrics. While some clustering methods are usually only implemented with specific distance metrics, the generalized algorithm was designed such that each clustering method would work with all 4 distance metrics. In order to assess the

effectiveness of each clustering method and distance metric we also develop software to generate simulated data sets containing known clusters. While this simulated microarray data cannot capture the all of the variety and complexity of experimental microarray data, it does enable investigation into the ability of hierarchical clustering algorithms to recover the know clusters contained in the data.

3.2 Materials and Methods

Simulated Data

4 sets of 50 simulated microarray data files were generated with known clusters and average within-cluster coefficients of variation of 0.15, 0.25, 0.35 and 0.45. Each file contained 500 genes with 10 ratio values per gene generated as log base 2 transformed ratios. To produce each cluster in a file, first a random number between 2 and 50 was generated to determine the number of genes for the cluster. Then the first ratio value for the first gene in the cluster was randomly generated based on a normal distribution with mean 0 and standard deviation of .5. The next ratio value for the first gene was randomly generated based on a normal distribution with a mean of the previously generated ratio value and standard deviation of .5. The remaining 8 ratio values were generated in a similar fashion to produce an expression profile for the gene consisting of 10 simulated microarray ratio measurements that tended to diverge from the initial mean of 0. Next a coefficient of variation, with a minimum value of 0.1, was randomly generated for each of the 10 ratio values to be used in producing subsequent gene expression profiles for the cluster. The range for these random numbers was determined by the desired within-cluster coefficient of variation for the data file. For

each subsequent gene in the cluster, each of the 10 ratio values was randomly generated based on a normal distribution using the corresponding ratio value from the initial gene profile as the mean and the corresponding coefficient of variation. An offset value that was randomly generated based on a normal distribution with mean 0 and a standard deviation of .25 was added to each expression profile, and each profile was multiplied by a scale factor that was randomly generated based on a normal distribution with mean 1 and a standard deviation of .25. Additional clusters were generated using this method until a total of 500 gene expression profiles were created, completing one data file. 50 data files were generated for each coefficient of variation (0.15, 0.25, 0.35, 0.45) for a total of 200 data files, each containing expression profiles for 500 genes, grouped into known clusters. The software to generate microarray data containing simulated clusters was developed using Microsoft Visual Basic due to its integrated graphical user interface design functionality. Figure 3.3 present the graphical user interface that is used to enter cluster parameters.

The screenshot shows a window titled "Form1" with the following fields and values:

Number of Files:	50		
Number of Variables:	10		
Number of Observations:	100		
Number of Clusters:	100	to	100
Observations per Cluster:	2	to	20
Cluster Variance	.1		
Total Variance	.25		
Replicate Variance	.25		
Replicates	10		

Buttons: "Generate" and "Calculate"

Figure 3.3 Input form for generating simulated microarray data for clustering

Input form for generating simulated microarray data organized into known clusters. Users can specify the number of samples (Variables), genes (Observations), and replicates (per gene), and ranges for the number of clusters per file, and the number of genes per cluster. Variances are specified between samples (Total Variance), genes within a cluster (Cluster Variance), and replicate genes (Replicate Variance).

Normalization and Centering

When processing microarray experimental data it is common to center and unit normalize data within each hybridization (column) for all genes, and across array experiments for each gene (row)(Eisen, 1998). Array centering and unit normalization is required to account for microarray experimental variation or bias. Gene centering and unit normalization adjusts for scale and offset differences between genes and allows them to be clustered based on the general shape of their expression profile, rather than absolute differences in their expression levels. Because the simulated expression ratios for this study were generated in assigned clusters, without experimental bias, it was undesirable to perform array (column) centering or unit normalization. However, each gene (row) was centered and unit normalized to account for the offset and scale that was introduced during data generation. This process involved centering each gene by subtracting the row mean and unit normalizing each gene by dividing by the magnitude of the row (Figure 3.4 and 3.5).

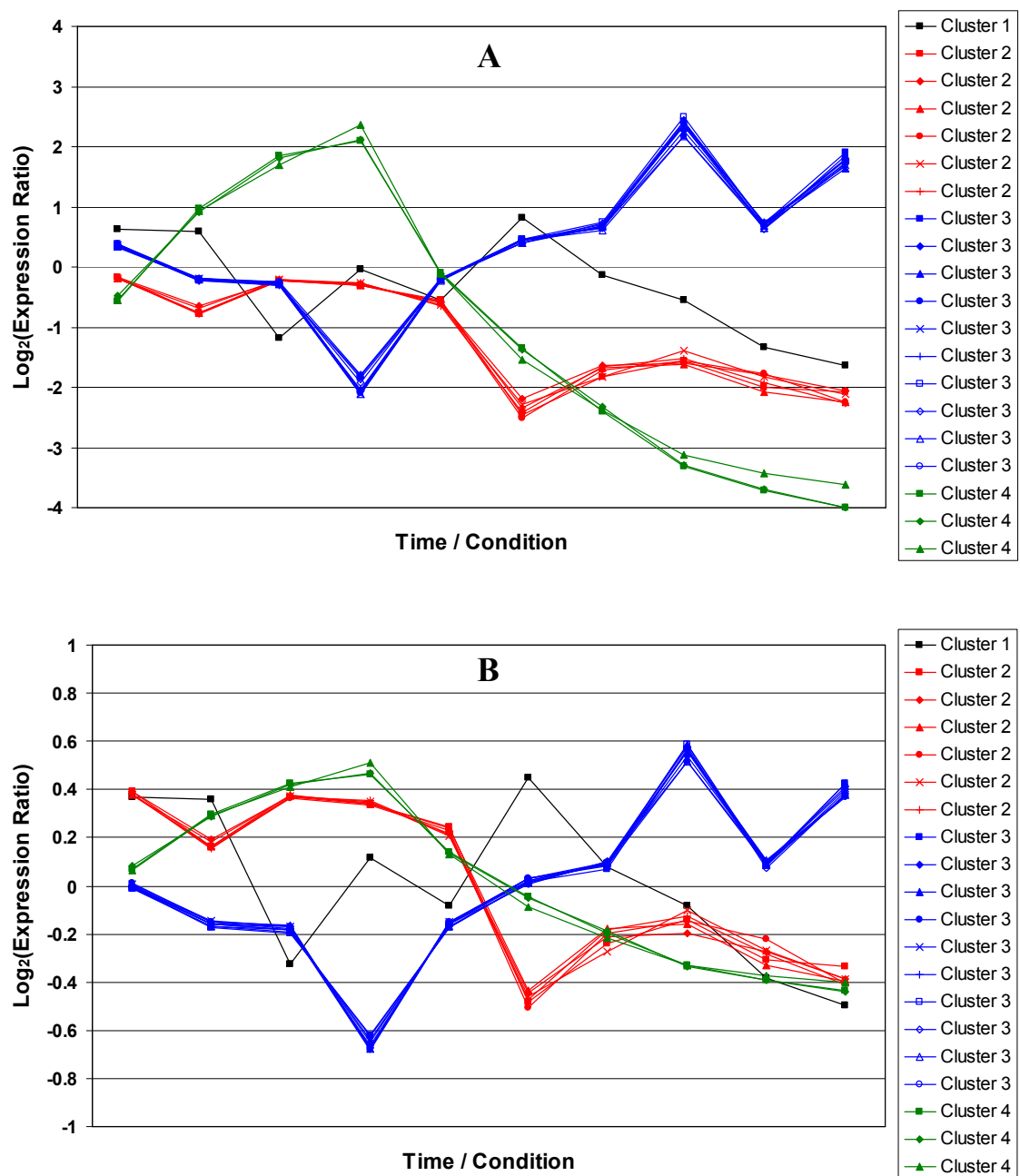


Figure 3.4 Simulated microarray datasets

Example of low cluster variance ($CV=10$) simulated dataset before (A) and after (B) mean centering and unit normalization. This data set contains simulated measurements for 20 genes at 10 time points contained within 4 clusters. Changes from A to B include a change in scale, centering around 0, and the normalization in scale of all clusters.

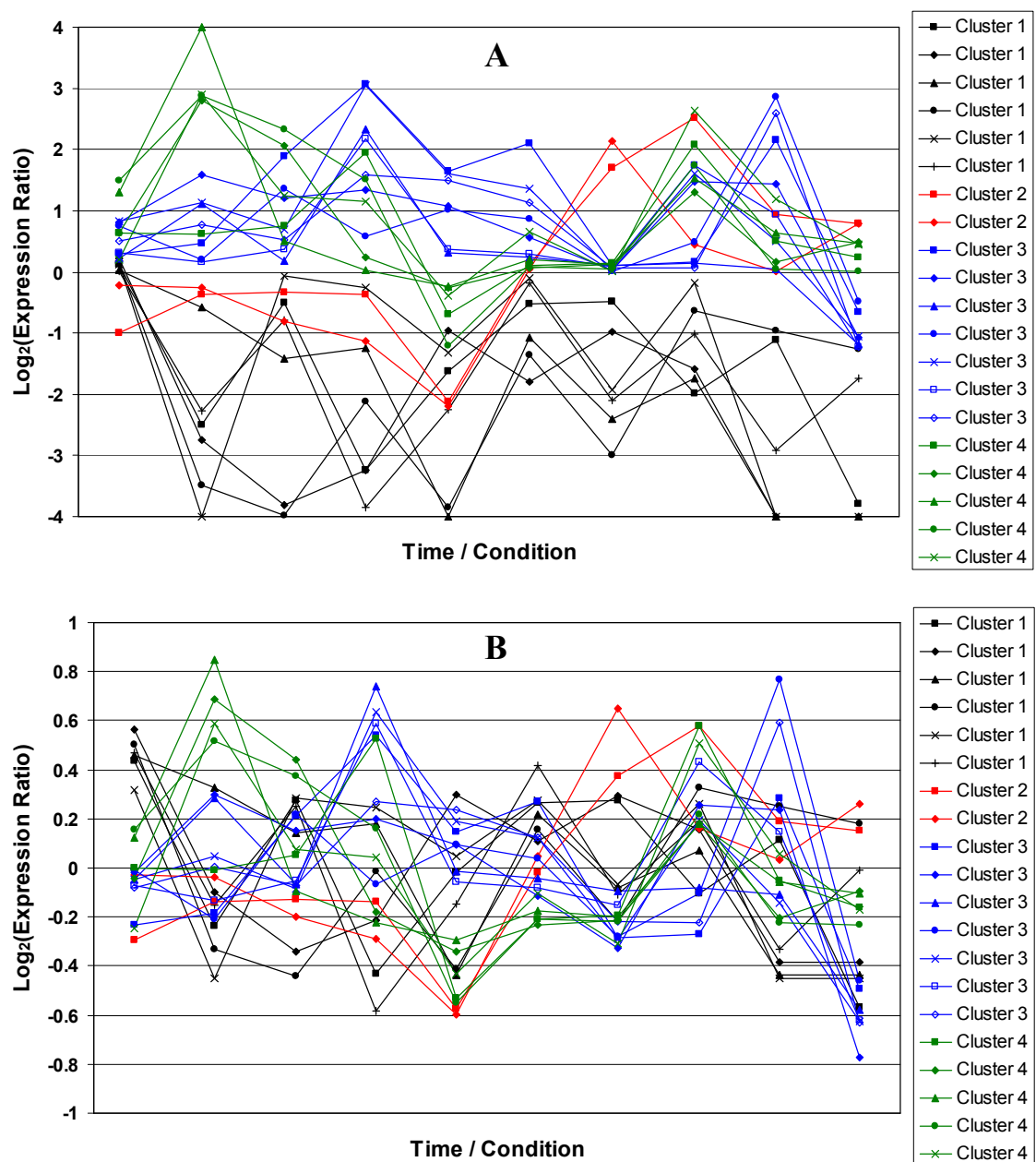


Figure 3.5 Mean centering and unit normalization

Example of high cluster variance ($CV=0.45$) simulated dataset before (A) and after (B) mean centering and unit normalization. This data set contains simulated measurements for 20 genes at 10 time points contained within 4 clusters. Changes from A to B include a change in scale, centering around 0, and the normalization in scale of all clusters. Also notice the obvious difficulty in distinguishing clusters as compared to the low variance charts graphs.

Hierarchical Clustering Algorithm

A generalized hierarchical clustering algorithm was developed to implement ten clustering techniques and four distance metrics. Specifics of the clustering techniques and distance metrics will be discussed in subsequent sections, however it is important to first understand the basics of the algorithm, which can be summarized as follows:

Step 1. Calculate and store the distance between every pair of items. An item refers to either a gene or a cluster; initially there are only genes.

Step 2. Create a cluster by joining the two closest items and flag the two items as clustered.

Step 3. Calculate and store the distance between this new cluster and all unclustered items.

Step 4. Repeat steps 1 and 2 until all items are clustered into one final cluster.

This clustering algorithm was implemented using the C programming language in a Microsoft Windows environment. A command line executable version of the program used in this study is available through the supplementary material accessible through the website. The program can be executed in either fast mode or memory saver mode. The fast algorithm, which stores the entire distance matrix in memory, executes rapidly for all hierarchical clustering methods but requires on the order of $(2n)^2$ bytes of memory where n represents the number of genes. Relatively small (<1000 genes) data sets will run efficiently on most personal computers, however memory requirements can be a limiting

factor for larger data sets. The memory saver mode alleviates this problem by storing, for each gene, only the index of the closest gene its calculated distance. Because the entire distance matrix does not reside in memory, the memory saver algorithm requires on the order of n bytes of memory; where n refers to the number of genes on the microarray. This modification allows essentially any size data set to be clustered on a computer with minimal memory, and has an acceptable impact on performance. For example, the fast clustering algorithm would require a minimum of 400 megabytes ($10000^2 * 4$ bytes) of memory to cluster a data set containing 5000 genes, where the memory saver algorithm would require a minimum of .04 megabytes ($5000 * 8$ Bytes) of memory.

Distance Metrics

The distance metric defines the formula used to calculate the distance between two items (an item refers to either a gene or a cluster). Distances are calculated using each of the $m=10$ corresponding pairs of ratio measurements for the two items being compared. When calculating a distance involving a cluster, the ratio measurements used in calculating the distance depends on the clustering method employed, which are discussed in the following section. The formulae for the distance metrics are listed below. In these formulae, i and j represent the indices of the two items being compared, and k represents the index of the ratio measurement. For this particular study all files contained $n=500$ genes and $m=10$ ratio measurements.

Euclidean Distance

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

Squared Euclidean Distance

$$d_{ij} = \sum_{k=1}^m (x_{ik} - x_{jk})^2$$

Manhattan Distance

$$d_{ij} = |x_{ik} - x_{jk}|$$

Distance based on Pearson Correlation Coefficient

$$d_{ij} = 1 - \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^m (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^m (x_{jk} - \bar{x}_j)^2}}$$

Hierarchical Clustering Methods

Hierarchical clustering methods differ in how the distance between clusters is calculated. The following list contains a brief description of each clustering methods. For a more detailed description please refer to references(Cormack, 1971)(Anderberg, 1973)(Sokal & Sneath, 1963).

Nearest Neighbor – Also known as single linkage, the distance between two clusters is defined by the distance between the two closest observations, one in each cluster.

Furthest Neighbor – Also known as complete linkage, the distance between two clusters is defined by the distance between the two most distant observations, one in each cluster.

Centroid – The distance between two clusters is defined by the distance between the midpoint (mean) of each cluster.

Median – Similar to the centroid method, the distance between two clusters is defined by the distance between the median point of each cluster.

Average – Also known as unweighted pair group method using arithmetic averages (UPGMA), the distance between two clusters is defined by the average distance between pairs of observations, one in each cluster.

Weighted Average – Also known as weighted pair group method using arithmetic averages (WPGMA) or McQuitty's similarity analysis, the distance between two clusters is defined by the total distance between pairs of observations, one in each cluster.

Total sum of squares – The distance between two clusters is defined by the total sum of squares (distance between each observation and the midpoint) of the cluster that would be created by joining the two clusters.

Average sum of squares – The distance between two clusters is defined by the average sum of squares of the cluster that would be created by joining the two clusters.

Incremental sum of squares – Also known as Ward's minimum variance, the distance between two clusters is defined by the total sum of squares of the cluster that would be created by joining the two clusters minus the sum of squares of each of the two individual clusters.

Flexible β - Similar to the weighted average method with the addition of a parameter β , which affects the shape of the resulting hierarchical tree. For this study β was set to -0.25.

Measure of Cluster Recovery

Data sets were generated with known clusters in order to evaluate the ability of each clustering method to recover these known clusters. While numerous potential indices of cluster recovery have been proposed (Rohlf, 1974), we used the Rand (Rand, 1971) statistic employed in the Milligan (Milligan, 1980) study. This statistic is based on

the n by n square matrix of entities that have been clustered, in this case genes. This statistic measures how well a given set of clusters matches the known clusters in the data. Because hierarchical clustering produces a hierarchical tree of clusters it was necessary to define the calculated clusters by selecting the clusters at the level of the hierarchy corresponding to the known number of clusters contained within the data. For this statistic, each entry δ_{ij} in the n by n gene matrix receives the value of 0 or 1. The value of δ_{ij} equals 1 if the genes corresponding to the row and column in the matrix were clustered together in the calculated clusters and in the assigned clusters. The value of δ_{ij} is also 1 if the genes corresponding to the row and column in the matrix were not clustered together in the calculated clusters and not in the assigned clusters. Otherwise δ_{ij} receives a value of 0. The final value of the statistic is between 0 (worst) and 1 (best) and is calculated as:

$$\frac{\sum_{i=1}^n \sum_{j>i}^n \delta_{ij}}{n(n-1)/2}$$

3.3 Results

Four sets of fifty simulated microarray data files were generated containing 500 gene expression profiles with known clusters and average within-cluster coefficients of variation of 0.15, 0.25, 0.35 and 0.45 (Figure 3.4a and 3.5a). All combinations of the 10 clustering methods and 4 distance metrics were employed to cluster the 500 gene expression profiles in each of the 200 simulated data files, following row centering and unit normalization (Figure 3.4b and 3.5b). The Rand statistic was calculated for each resulting hierarchical tree at the level of the hierarchy corresponding to the known

number of clusters contained within the data. The Rand statistic was averaged for the 50 files in each of the 4 file sets (Table 1).

Table 1 contains 4 groups of 4 columns each. Each group of 4 columns represents one of the 4 distance metrics: Euclidean, squared Euclidean, Manhattan, Pearson's correlation. Within each group are the four levels of within-cluster coefficients of variation of 0.15, 0.25, 0.35, 0.45. As expected, recovery of the known clusters within the simulated data sets diminished as the level of within-cluster variation increased. In addition, consistent with the findings of Milligan (Milligan 1980), the distance metric appeared to have little effect on cluster recovery, except for a small improvement at high levels of within-cluster variation when utilizing the Pearson's distance metric with the centroid and median methods. If row centering and unit normalization were not performed, only the distance metric based on Pearson's correlation coefficient still performed satisfactorily because the simulated expression profiles were generated with offset and scale factors (data not shown).

When looking at the cluster recovery of each clustering technique (row) in Table 3.1, the most striking feature is the comparatively poor performance of the nearest neighbor method, even at relatively low levels of within-cluster variance. While this method has some theoretical advantages (Fisher 1971) it has performed poorly in past studies comparing clustering procedures (Milligan, 1980) (Kuiper & Fisher, 1975), due to its tendency to incorrectly chain observations together. All other clustering techniques performed reasonably well, with the centroid and median methods exhibiting a greater degradation in cluster recovery as within-cluster variance increased.

The progression of each clustering method during the clustering process is presented in Figure 3.6. This figure contains graphs that plot the Rand statistic (red line), which measures cluster recovery, the R^2 statistic (blue line), which measures the degree to which the clusters explain the variance in the data, and the number of clusters consisting of at least 2 genes (black line). Each graph corresponds to one of the 10 clustering methods and presents these three metrics after each step in the clustering process. This clustering process proceeds from left to right, with no genes being clustered initially, and all genes belonging to the same cluster upon completion. This analysis was performed using the distance metric based on Pearson's correlation coefficient and a simulated data file containing nineteen clusters with an average coefficient of variation of .35.

Table 3.1 Comparison of cluster recovery

Comparison of cluster recovery for 10 agglomerative clustering algorithms in combination with each of 4 distance metrics for datasets with .15, .25, .35, and .45 coefficients of variation. Cluster recovery for each clustering algorithm and distance metric combination is the average Rand statistic for 50 semi-randomly generated datasets containing known clusters.

Clustering Method	Distance Metric / Coefficient of Variation															
	Squared Euclidean				Euclidean				Manhattan				Pearson's Correlation			
	0.15	0.25	0.35	0.45	0.15	0.25	0.35	0.45	0.15	0.25	0.35	0.45	0.15	0.25	0.35	0.45
Nearest Neighbor	0.93	0.63	0.23	0.15	0.93	0.63	0.23	0.15	0.92	0.61	0.21	0.15	0.93	0.63	0.23	0.15
Furthest Neighbor	0.98	0.95	0.92	0.92	0.98	0.95	0.92	0.92	0.98	0.95	0.93	0.92	0.98	0.95	0.92	0.92
Centroid	0.97	0.89	0.78	0.72	0.97	0.89	0.78	0.72	0.97	0.90	0.80	0.72	0.97	0.90	0.86	0.84
Median	0.97	0.88	0.80	0.72	0.97	0.88	0.80	0.72	0.97	0.90	0.80	0.72	0.97	0.91	0.88	0.87
Average	0.98	0.92	0.89	0.88	0.97	0.92	0.89	0.87	0.98	0.93	0.90	0.88	0.98	0.92	0.89	0.88
Weighted Average	0.97	0.93	0.91	0.90	0.98	0.93	0.91	0.90	0.97	0.94	0.92	0.90	0.97	0.93	0.91	0.90
Total Sum of Squares	0.99	0.97	0.95	0.93	0.99	0.97	0.95	0.93	0.99	0.97	0.95	0.93	0.99	0.97	0.95	0.93
Average Sum of Squares	0.98	0.95	0.93	0.92	0.98	0.96	0.93	0.92	0.98	0.96	0.94	0.92	0.98	0.95	0.93	0.92
Incremental Sum of Squares(Ward's Min Var)	1.00	0.98	0.96	0.94	0.99	0.98	0.96	0.94	0.99	0.98	0.96	0.94	1.00	0.98	0.95	0.94
Flexible β	1.00	0.97	0.95	0.93	1.00	0.98	0.96	0.94	1.00	0.98	0.96	0.94	1.00	0.97	0.95	0.93

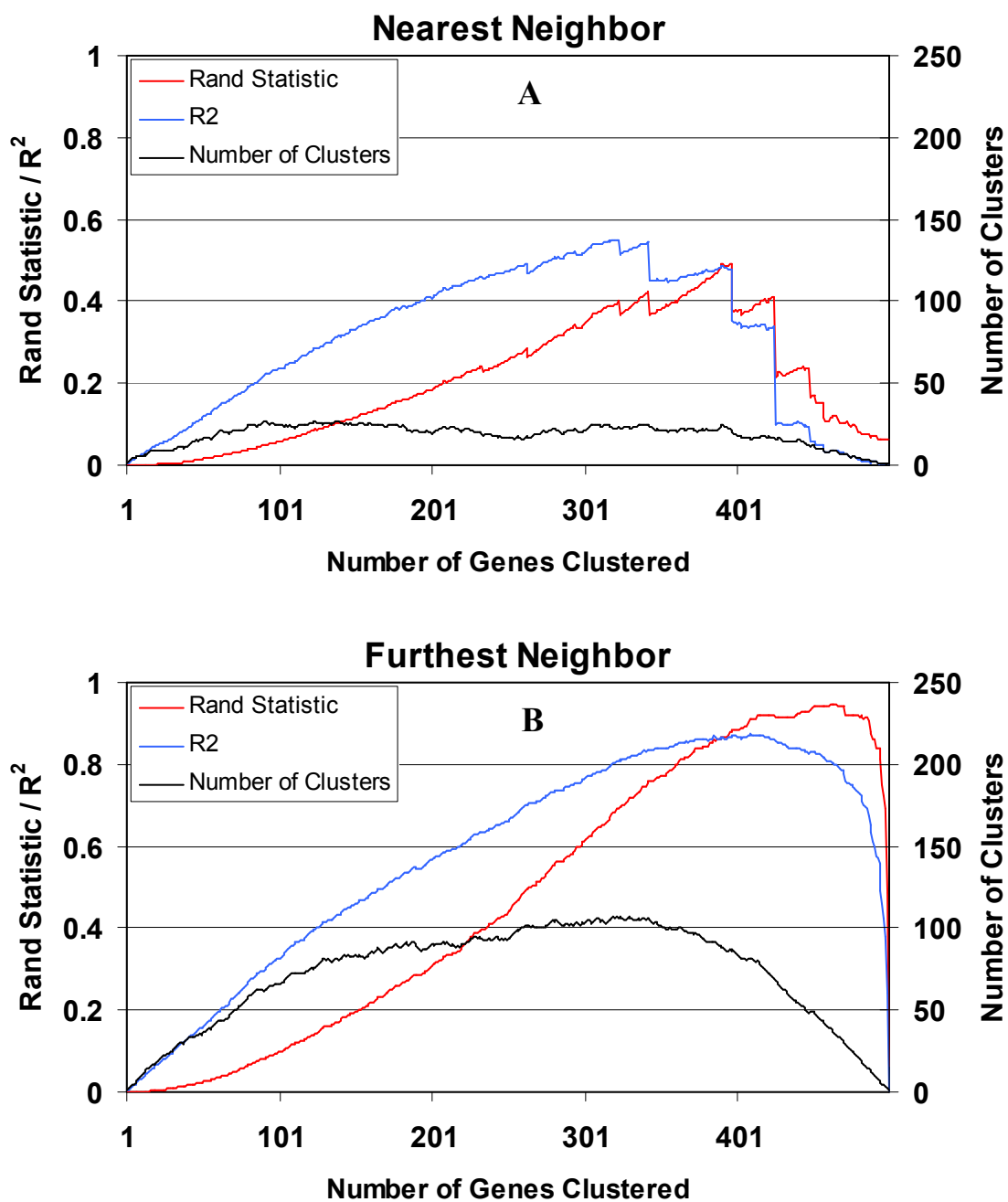


Figure 3.6 Progression of cluster recovery and cluster number for all algorithms
 The progression of the nearest neighbor (A) and furthest neighbor (B) clustering algorithms during the clustering process for an example dataset containing 500 genes. The Rand statistic (red line) measures cluster recovery, the R² statistic (blue line) measures the degree to which the clusters explain the variance in the data, and the number of clusters containing at least 2 genes (black line).

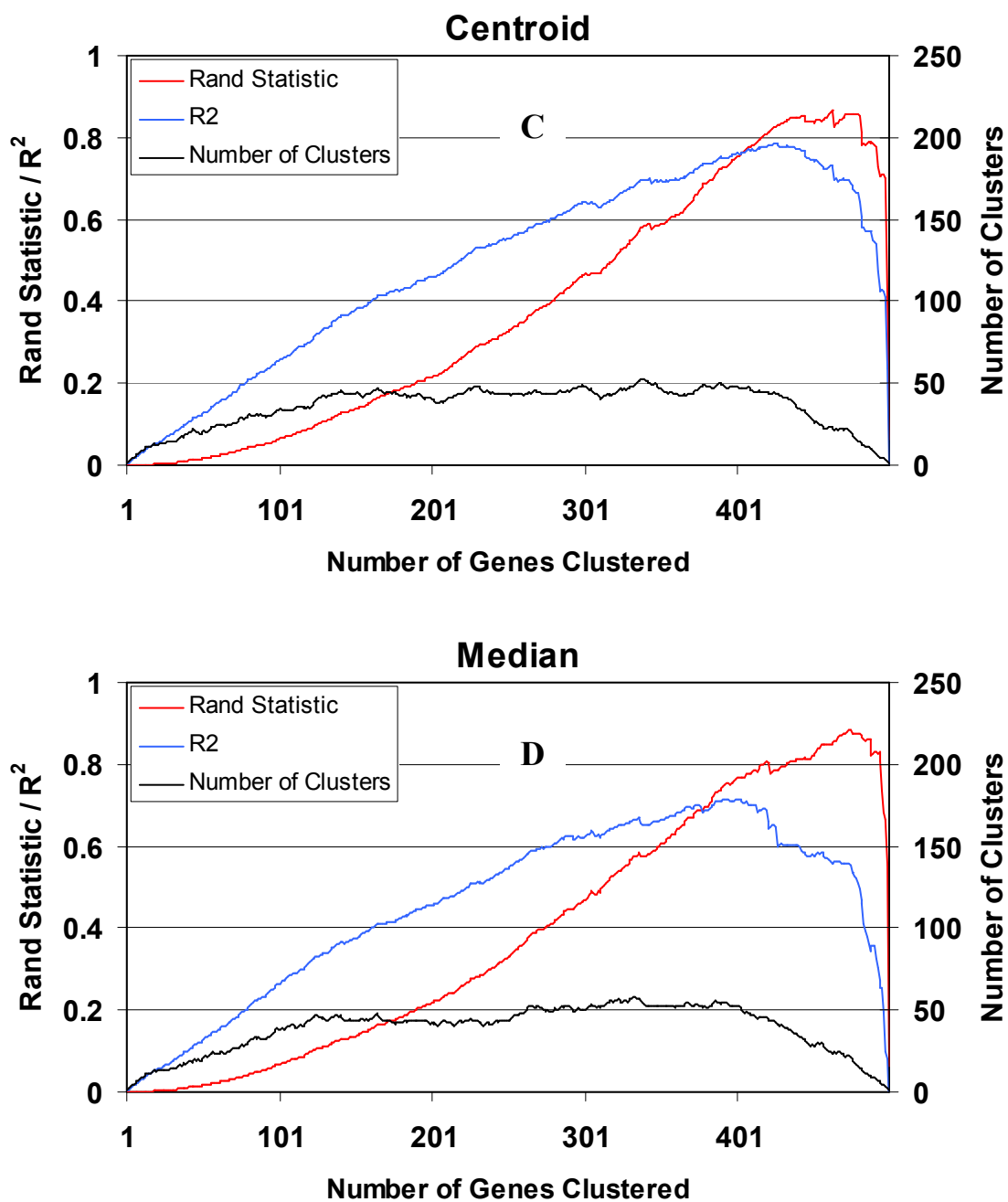


Figure 3.6 Continued

The progression of the centroid (C) and median (D) clustering algorithms during the clustering process for an example dataset containing 500 genes. The Rand statistic (red line) measures cluster recovery, the R^2 statistic (blue line) measures the degree to which the clusters explain the variance in the data, and the number of clusters containing at least 2 genes (black line).

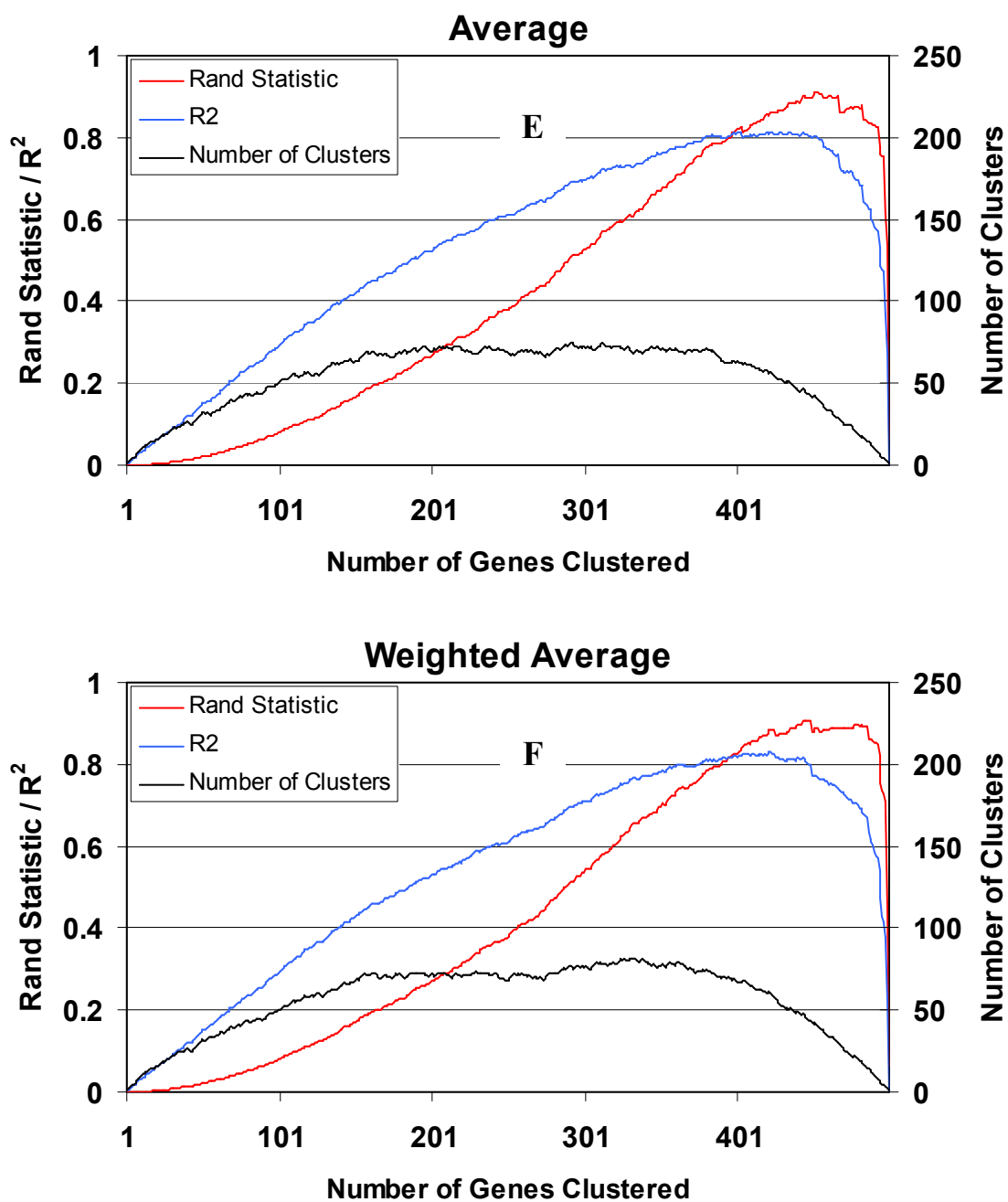


Figure 3.6 Continued

The progression of the average (**E**) and weighted average (**F**) clustering algorithms during the clustering process for an example dataset containing 500 genes. The Rand statistic (red line) measures cluster recovery, the R^2 statistic (blue line) measures the degree to which the clusters explain the variance in the data, and the number of clusters containing at least 2 genes (black line).

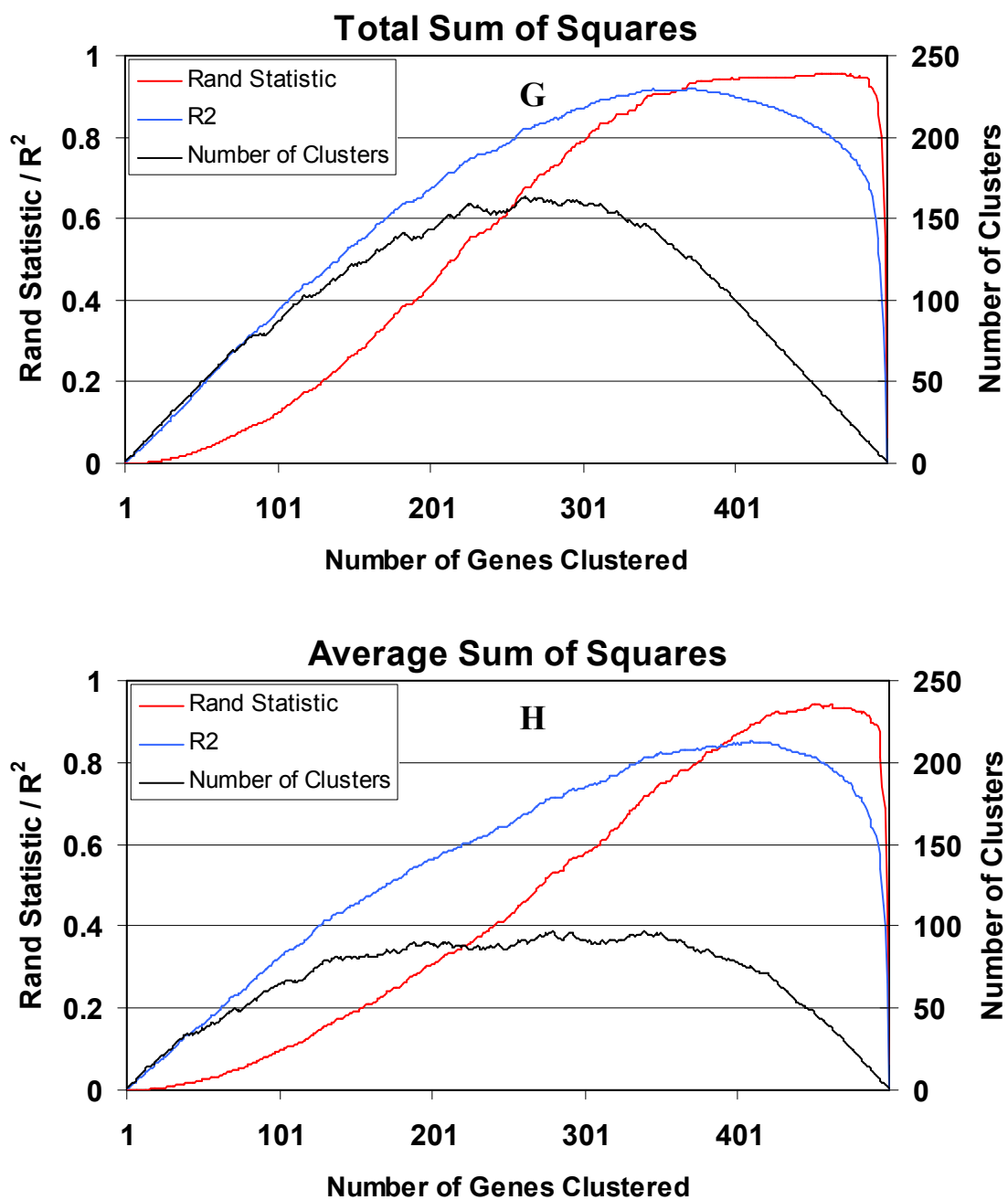


Figure 3.6 Continued

The progression of the total sum of squares (**G**) and average sum of squares (**H**) clustering algorithms during the clustering process for an example dataset containing 500 genes. The Rand statistic (red line) measures cluster recovery, the R^2 statistic (blue line) measures the degree to which the clusters explain the variance in the data, and the number of clusters containing at least 2 genes (black line).

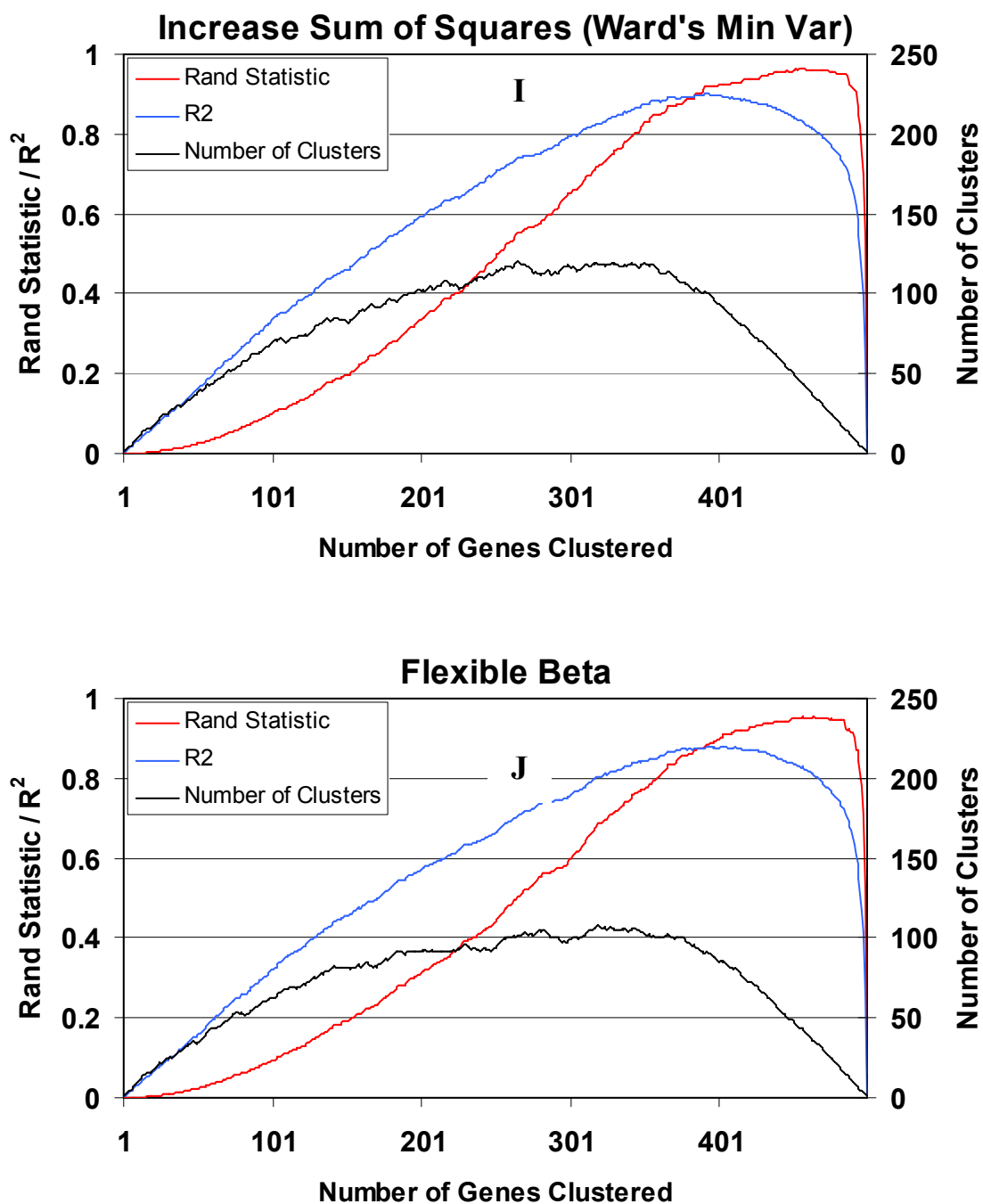


Figure 3.6 Continued

The progression of the increase sum of squares (**I**) and flexible beta (**J**) clustering algorithms during the clustering process for an example dataset containing 500 genes. The Rand statistic (red line) measures cluster recovery, the R² statistic (blue line) measures the degree to which the clusters explain the variance in the data, and the number of clusters containing at least 2 genes (black line).

3.4 Discussion

Cluster analysis was first described in the 1930's(Driver & Kroeber A.L., 1932), but did not receive significant attention until the early 1960's. This resurgence of interest was motivated by the availability of "high-speed" computers and a publication entitled *Principles of Numerical Taxonomy*(Sokal & Sneath, 1963), which advocated a radically empirical approach to biological taxonomy and clearly discussed a number of cluster analysis methods. This renewed interest in clustering stimulated the development of new clustering techniques in the 1960's and 1970's, and resulted in a number of published studies that evaluated the performance of many of these clustering algorithms on a variety datasets(Blashfield & Aldenderfer, 1978).

While we attempted to evaluate cluster performance based on simulated microarray data sets, Milligan reported on the effect of outliers and different types and amounts of error perturbation on cluster recovery(Milligan, 1980). With respect to outliers, the single linkage, centroid, and median methods proved robust, the average and weighted average demonstrated some susceptibility, and the flexible β and all sum of squares methods showed a marked decrease in performance. With respect to error perturbation, the average, weighted average, incremental sum of squares, average sum of squares, and flexible β performed well, followed by furthest neighbor. Nearest neighbor, centroid, median, and total sum of squares were all strongly affected by error perturbation. Kuiper and Fisher performed a comparison of six agglomerative hierarchical techniques and concluded that for compact clusters of nearly equal size, the increase sum of squares and furthest neighbor techniques performed best, while for

clusters that contained a varying numbers of observations, the centroid and average methods were preferable(Kuiper & Fisher, 1975).

This study was undertaken to investigate the applicability of various agglomerative hierarchical clustering techniques and distance metrics to microarray data. While this study cannot be used to infer the best hierarchical clustering technique for a specific data set, it does present the majority of agglomerative hierarchical clustering techniques and distance metrics that are applicable to microarray data. It is also important to note that most of these hierarchical clustering methods have some inherent bias. For example, the incremental sum of squares method is strongly biased towards producing clusters that contain the same number of observations due to its tendency to join clusters with a small number of observations. Complete linkage is strongly biased towards producing similar diameters because it joins clusters based on the distance between the most distant observations on each cluster. Therefore, it is important to select a clustering technique that is appropriate for the question of interest and the experimental design. The most appropriate clustering method for tissue classification may be different than the one that is most appropriate for grouping genes with similar expression profiles.

Whereas the choice of clustering technique often has more of an impact on clustering outcome than the choice of distance metric, the choice of distance metric can have a dramatic effect if gene (row) centering and unit normalization are not performed. In this case, the distance metric based on Pearson's correlation coefficient can produce considerably different results than the other three distance metrics. By looking at the formula for the Pearson's correlation distance metric (see Distance Metrics under

Materials and Methods) it is easy to see the reason for this discrepancy; mean centering and unit normalization is incorporated into the Pearson's correlation distance metric.

Therefore, it is important to consider the objective of hierarchical clustering when selecting the distance metric. If the offset and scale of the gene expression profiles should be taken into account when defining clusters, then the Pearson's correlation distance metric should not be used. If it is desirable to ignore the offset and scale of the gene expression profiles during clustering, then either Pearson's correlation distance metric should be used or gene centering and unit normalization should be performed.

It is also important to note that the applicability of any clustering method is dependent upon the data set being analyzed. The simulated data generated for this study contained known clusters that permitted evaluation of cluster recovery, however it is the purpose of clustering to define these clusters for experimental data. While there is no universally accepted method for evaluating cluster reliability, techniques involving introduction of random perturbation (Bittner *et al.*, 2000) and bootstrap resampling methods (Kerr & Churchill, 2001a) (Zhang & Zhao, 2000) have been utilized to augment cluster analysis and evaluate cluster stability.

4 GENE EXPRESSION ANALYSIS ON AN IN-VIVO MODEL OF ANGIOGENESIS AND BLOOD VESSEL MATURATION

4.1 Introduction

Angiogenesis is defined as the physiological process by which new blood vessels are formed from preexisting vessels(Ribatti *et al.*, 2004). Normal physiological processes that necessitate angiogenesis include embryonic development, ovulation, ischemic revascularization, and wound repair, and it is characterized by an eventual resolution into an organized hierarchical network of blood vessels consisting of arteries, arterioles, capillaries, venules, and veins(Dvorak, 2005). Angiogenesis is also evident in numerous pathological conditions such as proliferative diabetic retinopathies(Wegewitz *et al.*, 2005), solid tumor cancers(Folkman, 1985), rheumatoid arthritis(Luttun *et al.*, 2002), and chronic inflammatory diseases(Walsh & Pearson, 2001). This pathological angiogenesis is frequently characterized by chaotic, irregularly branched networks, that are structurally and functionally malformed, heterogeneously fenestrated, and are often permeable to plasma and plasma proteins(Dvorak, 2005). Inducing angiogenesis has shown promise in treating ischemic cardiovascular disease(Rosinberg *et al.*, 2004), peripheral vascular disease(Baumgartner *et al.*, 2005), non-healing ulcers(Heng *et al.*, 2000), and victims of strokes(Wei *et al.*, 2005). Understanding and affecting the mechanisms that control angiogenesis holds the promise of improving treatment for the major causes of mortality in the United States, including cardiovascular disease and cancer, which together account

for more than 50% of annual deaths(U.S. Department of Health and Human Services, 2005).

The physiological process of angiogenesis begins when endothelial cells in existing vessels become “activated” in response to some local stimuli including hypoxia, low pH, hypoglycemia, or mechanical factors such as flow or pressure(Pepper, 1997). As a consequence of this activation, portions of the vessels in the area near the stimuli become dilated, more permeable, and the parts of the basement membrane undergo proteolytic degradation, thus allowing plasma proteins to penetrate the extravascular space and form a fibrin gel. Endothelial cells then migrate from the existing vessel into the surrounding matrix to form a sprout, which continues to lengthen through endothelial cell division near the midsection and cell migration near the sprout tip. The endothelial cells that compose the sprout continue to replicate forming a tubular structure, which eventually inosculates with other sprouts or existing vessels to form a patent loop through which blood begins to flow. The sprouts that don't form permanent connections simply regress, however the new stable vessels complete their development through pericyte migration and recruitment and reconstitution of a basement membrane(Folkman, 1985)(Pepper, 1997). In addition, recent evidence suggests that circulating and marrow-derived endothelial precursor cells may contribute to adult angiogenesis in some circumstances(Rafii *et al.*, 2002).

At the molecular level, angiogenesis is thought to be controlled by both positive and negative regulators, which act in concert to produce a response to the local environment's blood flow requirements. The classic model, from which much of this

theory has been derived, is solid tumors. As early as 1945, researchers had demonstrated that tumor cells could elicit continuous growth of capillary endothelium(Algire & Chalkley, 1945), and by 1971 the first soluble factor that induced new capillary growth had been isolated(Folkman *et al.*, 1971). In 1972, researchers demonstrated that tumors that only grew to a small size on the anterior chamber of the eye where they could not become vascularized, grew exponentially when they were moved to a nearby location where they could attract vessels from the iris(Gimbrone, Jr. *et al.*, 1972). Factors that inhibited angiogenesis were first published in the early 1980's. Interferons(Brouty-Boye & Zetter, 1980), protamine(Taylor & Folkman, 1982), angiostatic steroids(Crum *et al.*, 1985), and thrombospondin-1(Rastinejad *et al.*, 1989)(Good *et al.*, 1990) have all been identified as angiogenesis inhibitors and therapies based on many of these agents have made it to clinical trials(Nature Publishing Group, 2005). Over the last 10 years numerous other endogenous inhibitors of angiogenesis have been discovered, many of which are fragments of naturally occurring extracellular matrix and basement membrane proteins (Table 4.1)(Cao, 2001)(Nyberg *et al.*, 2005).

Table 4.1 Endogenous inhibitors of angiogenesisAdapted from (Nyberg *et al.*, 2005)

Matrix Derived	Non-Matrix Derived
Arresten	Interferons
Constatin	Interleukins
Collagen Fragments	PEDF
EFC-XV	Platelet factor-4
Endorepellin	Angiostatin
Endostatin	Antithrombin III (cleaved)
Fibronectin Fragments	Chondromodulin
Fibulin	2-Methoxyestradiol
Thrombospondin 1 & 2	PEX
Tumstatin	Plasminogen Kringle 5
	Prolactin Fractions
	Prothrombin Kringle 2
	sFlt-1
	TIMP
	Troponin 1
	Vasostatin

The majority of our current understanding of the molecular mechanism regulating angiogenesis has only been developed within the last 15 years. By far the most widely studied class of angiogenic proteins are the vascular permeability factor/vascular endothelial growth factor family (VPF/VEGF) of proteins and their receptors. These proteins have been shown to participate in all types of vasculature development, including vasculogenesis, angiogenesis, and lymph-angiogenesis(Dvorak, 2005). Vascular endothelial growth factor (VEGF), originally called vascular permeability factor, was first discovered in 1983(Senger *et al.*, 1983), however it was not until 1993(Millauer *et al.*, 1993) that one of its receptors was identified and the groundwork for defining its pathway was laid. This founding member of the VPF/VEGF family, since renamed VEGF-A, can be found as multiple alternatively spliced isoforms, including the 206, 189, 165, and 121 amino acid forms. Its necessity for embryonic angiogenesis has been demonstrated by knockout mice generated through targeted disruption of VEGF-A and its two primary receptors VEGFR-1 (Flt-1) and VEGFR-2 (Flk-1)(Carmeliet *et al.*, 1996)(Ferrara *et al.*, 1996)(Fong *et al.*, 1995)(Shalaby *et al.*, 1995). VEGF-A expression is known to be regulated by a number of factors including hypoxia, nitric oxide (NO), cytokines and growth factors, various hormones, and oncogenes and tumor suppressor genes. As its original name suggests, most VEGF-A isoforms increase vascular permeability as well as increase endothelial proliferation, migration, and survival. VEGF-A alters endothelial cell gene expression by upregulating proteins associated with clotting and fibrinolysis, glucose transport, nitric oxide synthase, matrix metalloproteases, endothelial cell adhesion molecules, antiapoptotic factors, and various other transcription

factors(Dvorak, 2005). The major receptor responsible for communicating the effects VEGF-A is thought to be VEGFR-2, also known as Flk-1 and Kdr. The intracellular signaling pathway for VEGFR-2 is shown in Figure 4.1(Cross *et al.*, 2003). Numerous other growth factors and cytokines are now known to initiate or affect the process of angiogenesis. Examples of these molecules include fibroblast growth factors (FGFs), angiopoietins, transforming growth factor (TGF β), platelet-derived growth factors (PDGFs), tumor necrosis factor alpha (TNF- α), epidermal growth factor (EGF), interleukins, and angiogenin(Rundhaug, 2005).

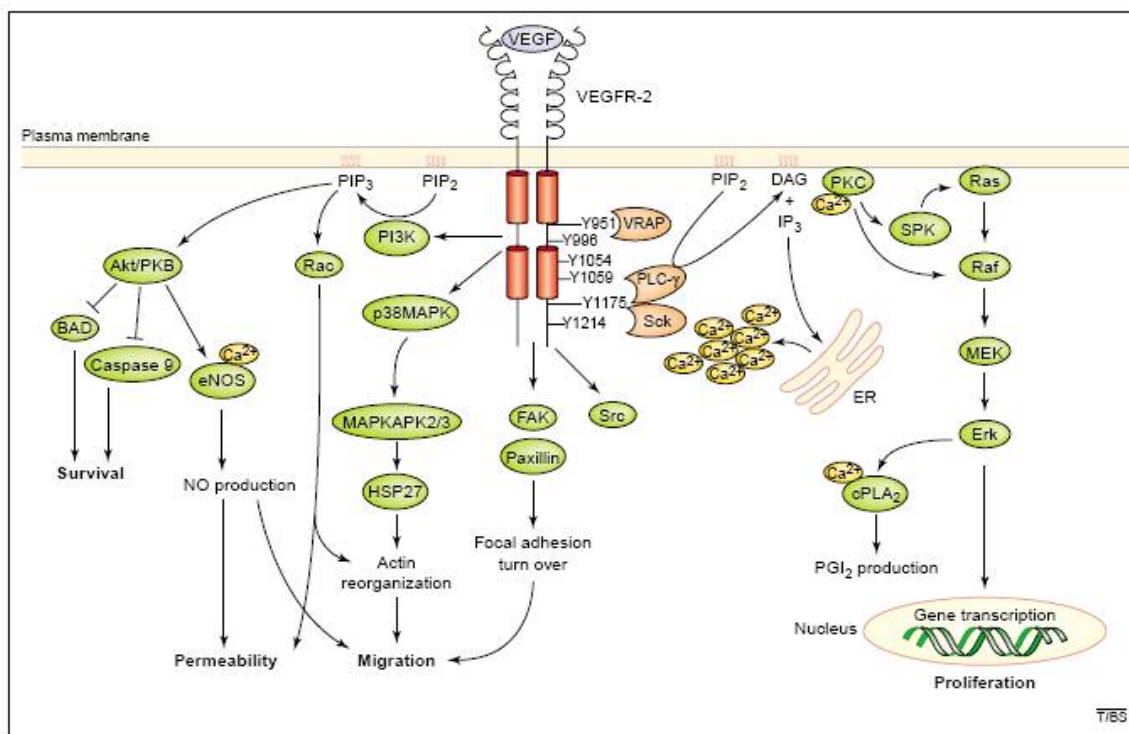


Figure 4.1 Vascular endothelial growth-factor receptor intracellular signaling

Schematic illustration of vascular endothelial growth-factor receptor (VEGFR-2)

intracellular signaling. Abbreviations: cPLA₂, cytosolic phospholipase A₂; eNOS, endothelial nitric oxide synthase; Erk, extracellular regulated kinase; HSP27, heatshock protein 27; MAPKAP 2/3, MAPK-activating protein kinase-2 and 3; NO, nitric oxide; PGI₂, prostacyclin; PIP₃, phosphatidylinositol (3,4,5)-trisphosphate; Sck, Shc-like protein; SPK, sphingosine kinase; VEGF, vascular endothelial growth factor (Cross *et al.*, 2003).

Both the basement membrane (BM) and extra cellular matrix (ECM) also influence the process of angiogenesis. Capillary basement membranes are primarily composed of collagens, perlecan, laminins, fibronectin, proteoglycans, nidogen/entactin, SPARC/BM-40/osteopontin, and nidogen/intactin as well as other molecules. This BM is thought to signal the endothelial cells to remain quiescent and to facilitate proper cell-cell adhesion(Form *et al.*, 1986). Researchers have also proposed a mechanism for endothelial cells to respond to mechanical forces such as sheer stress, changes in cell shape, or changes in connections to the BM or other cells. In the cellular tensional integrity (tensegrity) model, external forces are transferred across cell surface receptors the internal cytoskeleton through integrins and other adhesion receptors (eg. PECAM, E-selectin, cadherins), providing a local cell distortion-dependent mechanism of regulating growth, differentiation, motility, and apoptosis(Ingber, 2002). Cleavage fragments from a number of BM proteins have also demonstrated angiostatic properties, including Endostatin(O'Reilly *et al.*, 1997) and tumstatin(Maeshima *et al.*, 2000), which are fragments of collagen VIII and collagen IV respectively, and endorepellin(Mongiati *et al.*, 2003), which is a fragment of perlecan. Interestingly, these angiostatic peptides have all been shown to interact with integrins, however the functional receptor is different for each of them; $\alpha_5\beta_1$ for endostatin, $\alpha_v\beta_3$ for tumstatin, and $\alpha_2\beta_1$ for endorepellin, indicating possible different pathways for each peptide(Bix & Iozzo, 2005). Other matrix derived inhibitors of angiogenesis include arresten, which is derived from the α_1 chain type IV collagen, canstatin, which is derived from the α_2 chain of type IV collagen,

endostatin-like fragment, derived from type XV collagen, annastelin, a fragment of fibronectin, and fragments of fibulin 1D and domain III of fibulin 5(Nyberg *et al.*, 2005).

When disruption of the BM occurs, as in wounding, ECM bound angiogenic factors such as FGF are released (Arbiser, 1996) and coagulation causes platelets to adhere, amass, and degranulate releasing cytokines such as PDGF, TGF- β , VEGF, and interleukins(Kubota *et al.*, 2004)(Soslau *et al.*, 1997). Inflammation from wounding also recruits macrophages which release bFGF, TNF- α , and nitric oxide (NO), which in turn induces vasodilation and endothelial VEGF expression(Kimura *et al.*, 2000). VEGF increases vascular permeability and allows plasma proteins infiltrate the extracellular matrix thus establishing a temporary scaffold upon which activated endothelial cells can migrate(Conway *et al.*, 2001).

During most forms of angiogenesis the capillary BM is degraded by proteolytic enzymes, such as matrix metalloproteinases (MMPs). This degradation serves to dislodge the pericytes that support the vessel wall, and create an opening through which endothelial cells can migrate into the extracellular matrix. MMP activity also releases and activates many pro-angiogenic factors including: fibroblast growth factors (FGFs), transforming growth factors (TGFs), insulin like growth factors (IGFs), heparin-binding epidermal growth factor (HB-EGF), tumor necrosis factor alpha (TNF- α)(Rundhaug, 2005). MMP-9 cleaves the pro-angiogenic cytokine interleukin 8 (IL-8), increasing its activity ten-fold, in addition to degrading the angiogenesis inhibitor platelet factor-4(Van den Steen *et al.*, 2000).

Critical to the process of establishing a functional vasculature, whether through vasculogenesis or angiogenesis, is the recruitment and/or proliferation of vascular smooth muscle cells (VSMCs) and pericytes. Collectively known as mural cells, they stabilize new vessels by stimulating production of extracellular matrix and inhibiting endothelial cell proliferation and migration (Carmeliet, 2000), in addition to reducing vessel regression (Benjamin *et al.*, 1998). This mural cell covering provides hemostatic control, allowing the vasculature to respond to the changing needs in tissue perfusion, and assists the endothelial in performing specialized functions in various vascular beds (Hirschi & D'Amore, 1996). Endothelial cells produce and secrete a number of growth factors, including bFGF (Schultz & Grant, 1991) and PDGFB (Heldin & Westermark, 1999), which have been shown to stimulate mural cell proliferation and/or migration (Swinscoe & Carlson, 1992). In particular, mouse models deficient for PDGFB (Hellstrom *et al.*, 1999) or its mural cell receptor *pdgfrb* (Soriano, 1994) are embryonic lethal, and exhibit minimal blood vessel mural cell coverage. More recently an endothelium specific PDGFB knockout mouse has been developed, most of which survive into adulthood. Like the complete PDGFB knockout mice, these mice exhibit reduced vascular mural cell coverage and are prone hemorrhages, establishing the role of endothelial cell-derived PDGF as the source for pericyte recruitment (Bjarnegard *et al.*, 2004). Garmy-Susini *et al.* describes a possible mechanism for the close intercellular adhesion between endothelial cell and mural cells in neovasculature. Their work shows that integrin $\alpha 4\beta 1$ is expressed only in proliferating endothelial cells, while its ligand VCAM-1 is expressed only in proliferating mural cells. Antibodies against either integrin $\alpha 4\beta 1$ or VCAM-1

blocked the adhesion of mural cells to proliferating endothelial cells leading to apoptosis of both the endothelial cells and mural cells, thereby preventing neovascularization(Garmy-Susini *et al.*, 2005).

Both newly formed and existing vascular network demonstrate the ability to remodel by changing vessel diameter, type, distribution, or number to respond to a variety local environmental stimuli. Because of the complexity of the process of vascular development and adaptation most researchers have focused on individual components of the process including vasculogenesis, angiogenesis, intussusception, regression, and arteriogenesis, and vein/venule remodeling(Peirce & Skalak, 2003). Vasculogenesis is the process of forming new vessels *de novo*, and although it is usually used to describe blood vessel formation during embryogenesis it has recently been implicated in adult capillary formation(Murasawa & Asahara, 2005). Intussusception, often classified as an alternative form of angiogenesis, is the process by which the lumen of an already perfused vessel is partitioned into two or more compartments through the ingrowth of interstitial tissue structures or tissue posts, followed by network expansion(Patan *et al.*, 2001)(Patan *et al.*, 1992). Vascular regression is the selective degeneration of blood vessels within a vascular network. The antithesis to angiogenesis, this process occurs during embryonic development(Watanabe *et al.*, 2001) as well as in response to many of the same stimuli including hyperoxia(Alon *et al.*, 1995), changes in various cytokines and growth factors(Thurston *et al.*, 2005), metabolic demand, extracellular matrix(Jang *et al.*, 1998), and blood flow and pressure, as well as lack of perivascular cell contact(Benjamin *et al.*, 1999)(Peirce & Skalak, 2003). Arteriogenesis refers to two different mechanisms

by which arteries or arterioles forming from existing vessels in response to local stimuli, inflammation, and hemodynamic stress(Scholz *et al.*, 2001). Arterialization is the process by which capillaries recruit perivascular cells, which differentiate into smooth muscle cells, to form arterioles(Price & Skalak, 1996). Collateralization is the process by which collateral vessels enlarge through endothelial cell activation, vascular cell proliferation, and matrix remodeling(Buschmann & Schaper, 2000). In addition to remodeling the inflow vessel (arteries and arterioles), a vascular network must be able to remodel the outflow vessels (veins and venules). While hemodynamics has certainly demonstrated a role in determining the vein/venule phenotype the molecular mechanisms are much less well characterized. Possible molecules involved in the process include angiopoietin-1, Tie2, Eph-B4, P-selectin, and E-selectin, however much of the molecular mechanism remain to be uncovered(Thurston *et al.*, 2000).

Mathematical models are another method by which researchers are investigating the complexities of vascular adaptation. While the first mathematical model proposed over 80 years ago was based primarily on vascular wall shear stress(Murray, 1926), researchers have now demonstrated that functional vascular remodeling requires the interpretation of several stimuli including shear stress, circumferential stress, metabolic status of the tissue, and the propagation of stimuli upstream and downstream along vascular segments(Pries & Secomb, 2005). Researchers have also added cell level stimuli including epigenetic signals, molecular signaling, and cellular behaviors to their models in an effort to improve microvascular pattern prediction(Peirce *et al.*, 2004). The

impact of the extracellular matrix has been added to some models and is helping to shed light on its influence on angiogenesis and vascular patterning(Sun *et al.*, 2005).

All mathematical models of vascular adaptation require experimental measurements of the parameters that influence the implementation of the model. These parameters include physiological measurements such as vessel segment lengths and diameters(Gruionu *et al.*, 2005), blood flow and pressure, and the effects blood rheology on the apparent viscosity as a function of blood vessel diameter and bifurcations(Pries *et al.*, 1990)(Pries *et al.*, 1994)(Pries & Secomb, 2005), as well as experimental evidence of the molecular signals that influence that vascular architecture(Serini *et al.*, 2003). Mathematical models provide a mechanism to integrate and investigate our current understanding of angiogenesis and vascular remodeling, however to realize their full potential they must be developed based on sound experimental evidence.

The process of angiogenesis, blood vessel maturation, and vascular network remodeling is not only critical to embryonic development, but also a significant factor in the progression of many diseases. Numerous local environmental factors including sheer stress, circumferential stress, metabolic state, inflammation, and endocrine, paracrine, autocrine and propagated signals are continuously instructing the vasculature to adapt to the requirements of both the local and regional tissue as well as the cardiovascular system(Pries & Secomb, 2005). All of these stimuli, however, must at some point be translated through a cellular mechanism to affect changes in the vasculature. For example, hypoxia increases hypoxia-inducible factor-1 α (HIF-1 α) protein levels through decreased ubiquitination and degradation, which in-turn form heterodimers with HIF-1 β

subunits and affect the transcription of many angiogenesis related genes including VEGFA (Hewitson & Schofield, 2004). These cellular mechanisms involve numerous forms of cell signaling (i.e. changes in transcription, translation, phosphorylation, glycosylation, Ca⁺ signaling, etc.), however techniques do not yet exist to quantitate most of these mechanisms on a large scale. Fortunately, many of these events alter mRNA levels through changes in transcription or degradation, which can be measured using microarrays.

In an effort to improve our understanding of the cellular mechanisms regulating angiogenesis, blood vessel maturation, and vascular remodeling we utilized a mouse microvessel fragment model to study gene expression during the formation of a vascular network from small vessel fragments isolated from mouse periovarial and epididymal fat pads. Following isolation, microvessel fragments were embedded in type I collagen, which was pH neutralized and allowed to gel, and implanted subcutaneously on the hindquarters of severe combined immunodeficient (SCID) mice. During the first week after implantation these microvessel fragments underwent sprouting angiogenesis to form networks of small diameter vessels. Inosculation with the host vasculature occurs near the end of the first week, and although initial blood flow is atypical it resolves to a near physiological pattern over the following three weeks. Over the course of these following three weeks these networks remodeled into a typical vascular network consisting of inflow vessels (arteries and arteriols), capillaries, and outflow vessels (veins and venules). Total RNA was extracted from implants explanted on days 3, 7, 14, 21, 28, as well as freshly isolated vessels (day 0). The total RNA was then amplified, fluorescently

labeled, and comparatively hybridized between time-points to determine changes in gene expression during the formation of the new vasculature. By defining patterns of gene expression and clustering differentially expressed genes based on expression profile, we have begun to uncover the genetic mechanism that regulates progression through the vascularization process.

4.2 Results

Microvessel fragments were isolated from the periovarial and epididymal fat pads of TIE2-GFP mice, whose vascular endothelial cells express GFP under control of the Tie2 (also known as Tek) promoter (Motoike *et al.*, 2000). Isolated adipose tissue was minced and partially digested with collagenase in order to isolate small microvessels consisting of fluorescent endothelial cells and non-fluorescent mural cells (pericytes and smooth muscle cells). These isolated vessels were combined with type I rat tail collagen (3 mg/ml) and delbecco's modified eagle media (1 x final concentration) at a density of 15,000 vessels per milliliter, pH neutralized, and placed in an incubator at 37⁰ C for 20 minutes to facilitate polymerization. Polymerized gels were then implanted into the hindquarters of non-transgenic severe combined immunodeficient (SCID) mice. During the first week of implantation many of the isolated vessels underwent sprouting and elongation to form a somewhat homogeneous meshwork of small diameter vessels (Figure 4.2). Vessels within the gel construct retain some α -actin positive perivascular cells, however they are distributed sparsely throughout the new vascular network, and some of the perivascular cells seem to have disassociated from vessels (Figure 4.3). During the second week some of the small diameter vessels appear to remodel into larger

caliber vessels (Figure 4.2), however the overall topology of the vascular network is still highly irregular. In particular, these larger caliber vessels exhibit variable diameters and unconventional branching. A primitive vascular network appears to be developing (Figure 4.2), however there is still a disproportionate percentage of the small diameter vessels present at day 14. By the end of the third week post-implantation (Figure 4.2), the vascular network displays many of the same characteristics of a normal physiological vascular bed. Most of the vessels present a uniform diameter, and the expected changes in vessel diameter are present at branching sites. 28 days post implantation the new vascular network is almost indistinguishable from a normal physiological vascular bed. Both network topology (Figure 4.2) and mural cell coverage (Figure 4.3) indicate the presence of all of the vessel types (arteries, arterioles, capillaries, venules, and veins) present in a normal vasculature.

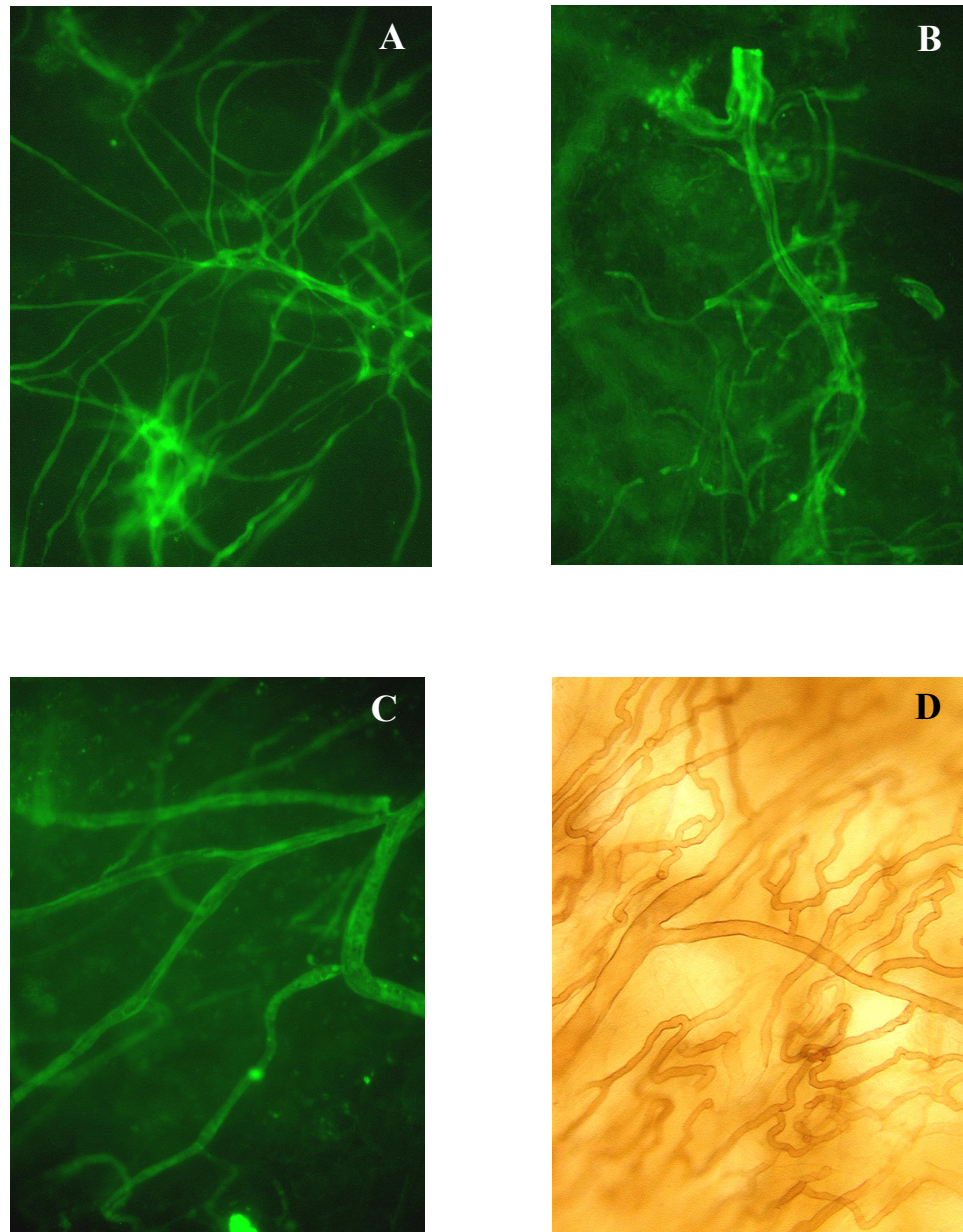


Figure 4.2 Images of microvessel implants explanted at days 7, 14, 21, 28
Confocal microscopy images of microvessel implants that were explanted at days 7 (A), 14 (B), 21 (C), and 28 (D). Panels A, B, and C are fluorescence images of GFP vessels. Panel D is a GS-1 stained image. Courtesy of Helen Chen.

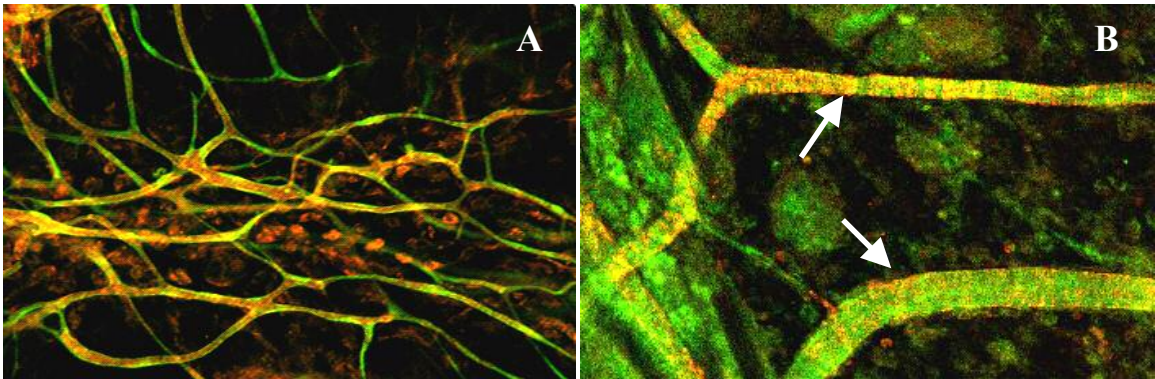


Figure 4.3 Mouse microvessel explants stained for smooth muscle actin

Fluorescence confocal images of microvessel implants that were explanted at days 7 (**A**) and 21 (**B**) and stained for smooth muscle actin (conjugated with Cy5). GFP fluorescence is pseudo colored green and Cy5 fluorescence is pseudo colored red. Vessels in panel **A** exhibit sporadic mural cell coverage with some of the mural cells appearing to be disassociated from the vessels. Vessels in panel **B** display the more characteristic mural cell coverage associated with arteries (top arrow) and veins (bottom arrow). Courtesy of Helen Chen

As another measure of vascular development, blood perfusion was assessed via intravascular injection of rhodamine-conjugated dextran into the host SCID mouse. Blood perfusion was assessed using both real-time intravital video imaging and confocal microscopy. Imaging was performed using the appropriate wavelengths for GFP, in addition to rhodamine, in order to distinguish between the host vasculature and vessels derived from the implanted microvessel fragments. Examination of blood perfusion at the end of the 1st week revealed that while vessels within the construct had inosculated with the host vasculature, rhodamine fluorescence was limited to the periphery of the implant (Figure 4.4). By the end of week two many of the GFP-positive vessels within the construct exhibited rhodamine fluorescence, however the perfusion pathways were often irregular and inefficient. As an example, in one instance blood flowed from a smaller caliber vessel into a larger diameter vessel which then branched into smaller diameter vessels (Figure 4.4). By week four however, blood flow seemed to resolve into a normal physiological pattern, in which blood flows from larger diameter vessels (arteries), which branch into arteriols that feed capillaries that drain into venules, which terminate at the final vein outflow vessel (Figure 4.4). Confocal images of taken after rhodamine-conjugated dextran injection are also indicative of an immature vasculature at day 7 and a more mature vascular at day 21, which contains vessels characteristic of both arteries and veins (Figure 4.5).

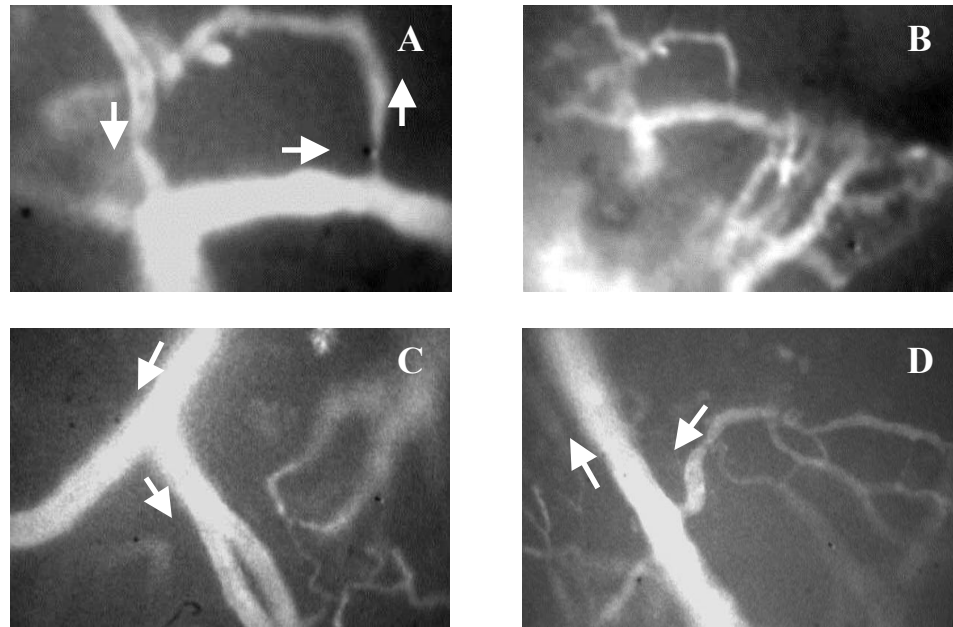


Figure 4.4 Intravital images of blood flow in microvascular implants

Images from real-time intravital video after intravascular injection of rhodamine-conjugated dextran into the host SCID mouse after 14 days (**A** and **B**) and 28 days (**C** and **D**) of implantation. Arrows indicate the direction of observed flow. At day 14 blood flow was observed to be abnormal, with flow in some segments initiating at a large vessel in terminating at a relatively small vessel (**B**). By day 28 flow has resolved into a normal pattern with larger vessels branching into smaller vessels (**C**), and smaller vessels draining into larger vessels (**D**). Courtesy of Chris Sullivan.

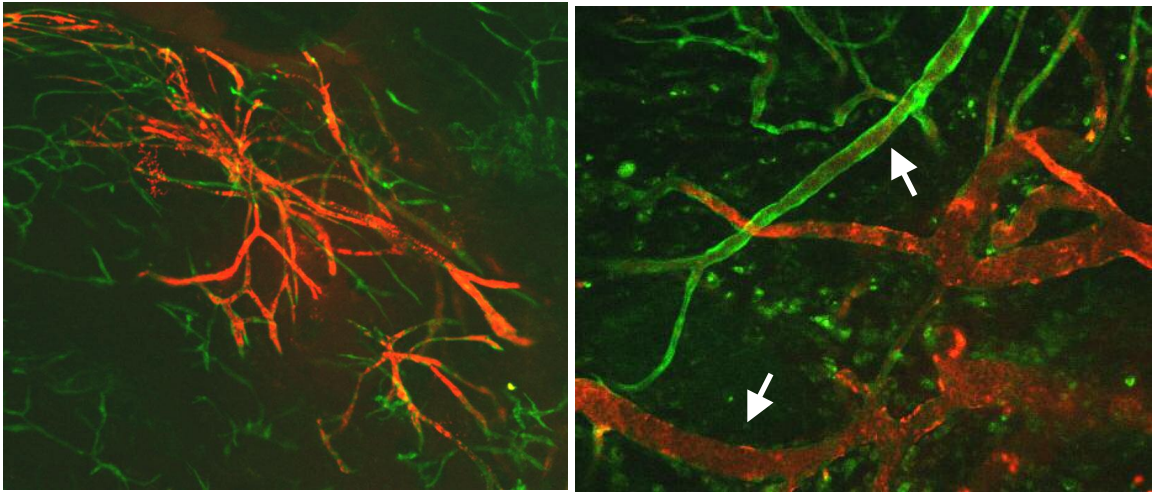


Figure 4.5 Images of implants after intravascular injection of rhodamine dextran

Fluorescence confocal images after intravascular injection of rhodamine-conjugated dextran into the host SCID mouse after 7 days (**A**) and 21 days (**B**) of implantation. GFP fluorescence is pseudo colored green and Cy5 fluorescence is pseudo colored red. At day 7 (**A**) only some of the vessels are perfused and vessel morphology is indicative of immature vasculature. By day 21 (**B**) most of the vasculature is perfused and some vessels appear to have matured into arteries (top arrow) and veins (bottom arrow).

Courtesy of Chris Sullivan.

Proliferation in the constructs was assessed using BrdU incorporation (Figure 4.6). These experiments revealed that vascular cell proliferation peaked during the 1st week of implantation and subsided to a lower level by day 14. Proliferation reached its lowest level by day 21, and appeared to increase slightly at day 28 indicating that network remodeling had not subsided entirely. During the 1st week of implantation proliferation was present somewhat uniformly throughout the construct, however at later weeks BrdU incorporation was limited primarily to larger vessels. These measurements are consistent with the observed network topology in that some of the smaller vessels present at day 7 appear to have remodeled into larger vessels by day 28. For this process to occur, cells within the vascular wall must proliferate in order to expand the vessel diameter.

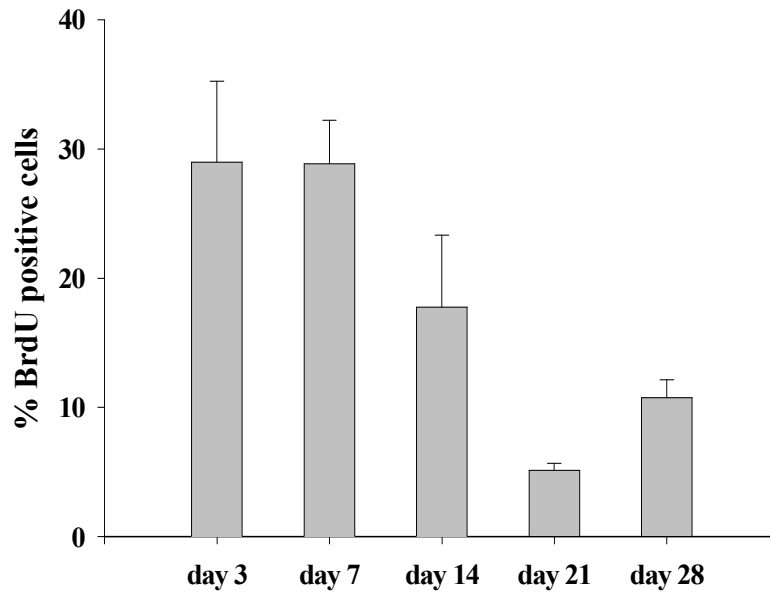


Figure 4.6 Cell proliferation at 3, 7, 14, 21, and 28 days post implantation

Cell count of positively stained cells after BrdU incorporation. Cell proliferation is maximal at days three and seven and significantly subsides by day 21. Proliferation at later stages was limited primarily to larger caliber vessels (data not shown). Courtesy of Helen Chen.

These morphological and molecular measurements of the implanted microvessel constructs provide sound evidence that a functional vascular network can develop from small microvessel fragments in the relative absence of non-vascular cells. In addition, this process occurs in a highly organized fashion, with a highly angiogenic phase during the first week being followed by a process of vessel maturation and network remodeling during the following three weeks. Because this model of vascularization is relatively free of other cell types it makes it ideal for microarray experiments, which measure the combined expression level of all cells in a sample. In an effort to measure changes in gene expression that occur during this process of vascular development, total RNA was extracted from microvessel constructs that were explanted at days 3, 7, 14, 21, and 28 as well as freshly isolated vessels. Total RNA was then amplified in order to obtain sufficient RNA to perform four hybridizations per sample. Amplified RNA samples were reverse transcribed and coupled with either the Alexa Fluor 546 or Alexa Fluor 647 fluorochrome and hybridized to a custom mouse cDNA microarray containing the national institutes of aging (NIA) 15,000 cDNA cloneset (Kargul *et al.*, 2001), using an interwoven loop experimental design (Figure 4.7) (Kerr & Churchill, 2001b). This entire process was actually performed two separate times, with different vessel source and host animals. The experimental design differed slightly between the two replicates, with the day 21 sample being added to the experimental design before the second experiment (Figure 4.7). A day 21 sample was later added to the design of the first experiment.

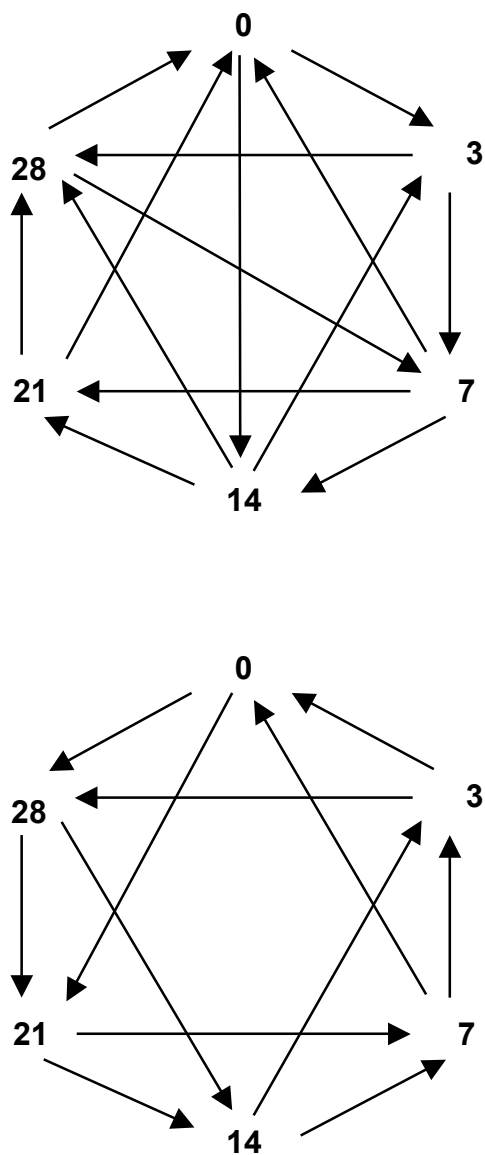


Figure 4.7 Hybridization scheme for mouse microvessel experiment

Hybridization scheme for each replicate of the time course experiment. Each time point is indicated at the corners of the diagram and arrows represent each hybridization that was performed. The tail of the arrow represents labeling with the Alexa Fluor 546 fluorochrome and the head of the arrow indicates labeling with the Alexa Fluor 647 fluorochrome. In the first hybridization scheme (A) day 21 was added later, which is the reason why there are additional hybridizations as compared to the second hybridization scheme (B). Note also that the arrows common to both panels are reversed in panel B as compared to panel A to offset any dye bias.

The resulting dataset was analyzed using a gene-by-gene ANOVA that included terms for array, dye, gene, and time-point effects, performed with the software (CARMA) that was developed as part of aim 1. Only genes that were consistently confidently measured (background subtracted intensity $> 2 * \text{local background standard deviation}$) for at least one sample were used in the ANOVA, however all spots were used for pre-processing of the dataset. Genes were identified as differentially expressed based on a shrunk ANOVA p-value based on the James-Stein shrinkage concept (Cui *et al.*, 2005) of less than or equal to .05 for the time-point term after FDR adjustment (Benjamini & Hochberg, 1995). Out of the 3,470 genes analyzed, 444 genes were identified as differentially expressed. As part of the ANOVA, estimates of differences in gene expression between each of the explanted samples and the day 0 sample were generated. These values were then mean centered and unit normalized and then clustered using the Ward's minimum variance agglomerative hierarchical clustering algorithm identified as part of aim 2. Based on the hierarchical tree generated by the agglomerative clustering process, 14 clusters were identified (Figure 4.8). The genes contained within each cluster are presented in Appendix 2. Real-time PCR was performed for a select group of genes, some of which were present on the microarray, and some of which were identified as possible genes of interest based on suspected pathways and review of the literature (Table 4.2).

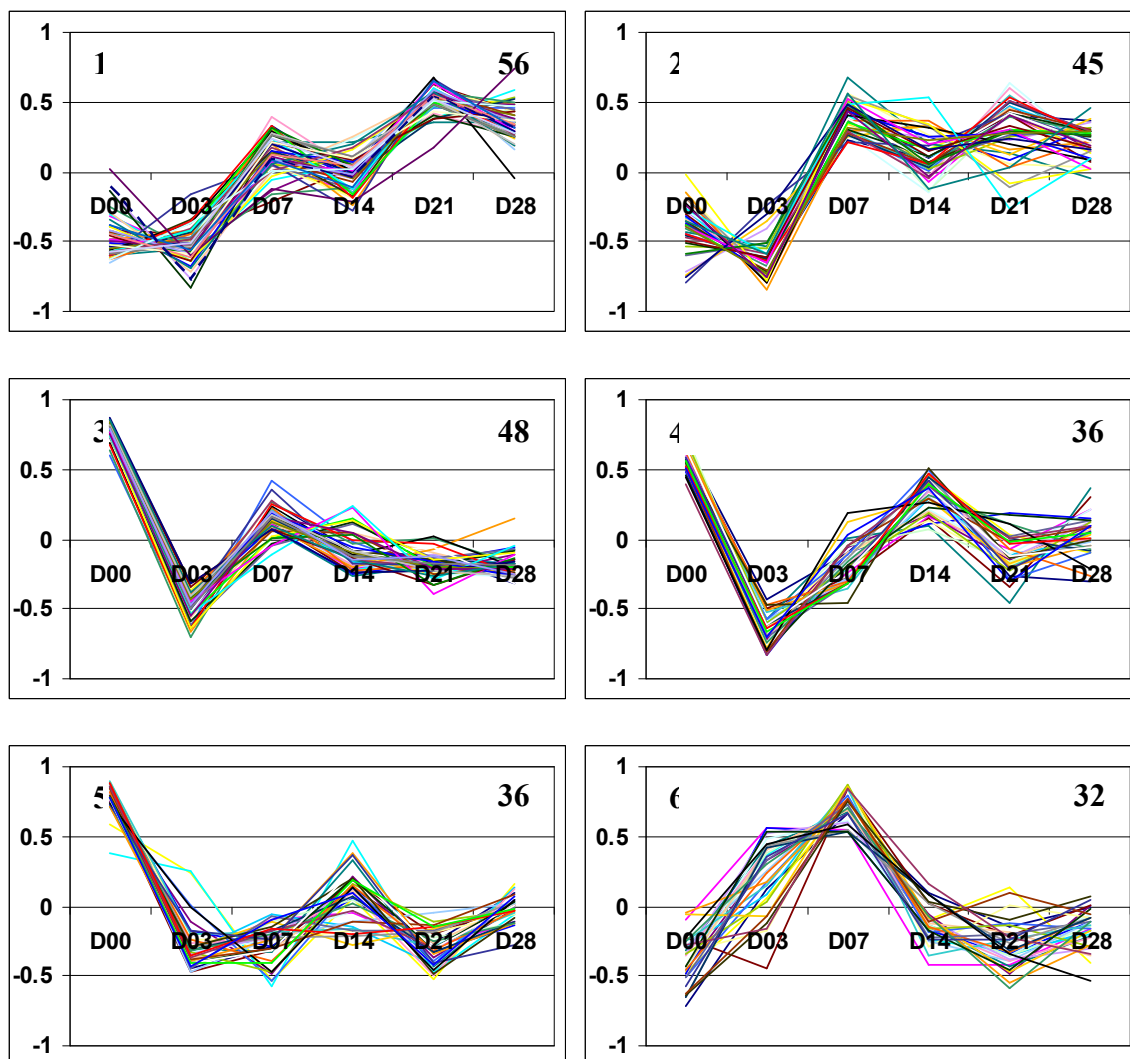


Figure 4.8 Mouse microvessel gene expression clusters for timecourse experiment
 The 14 clusters generated by performing increase sum of squares agglomerative hierarchical clustering on the 444 genes identified as differentially expressed. The number of genes contained within the cluster is indicated in the upper right hand corner of each chart.

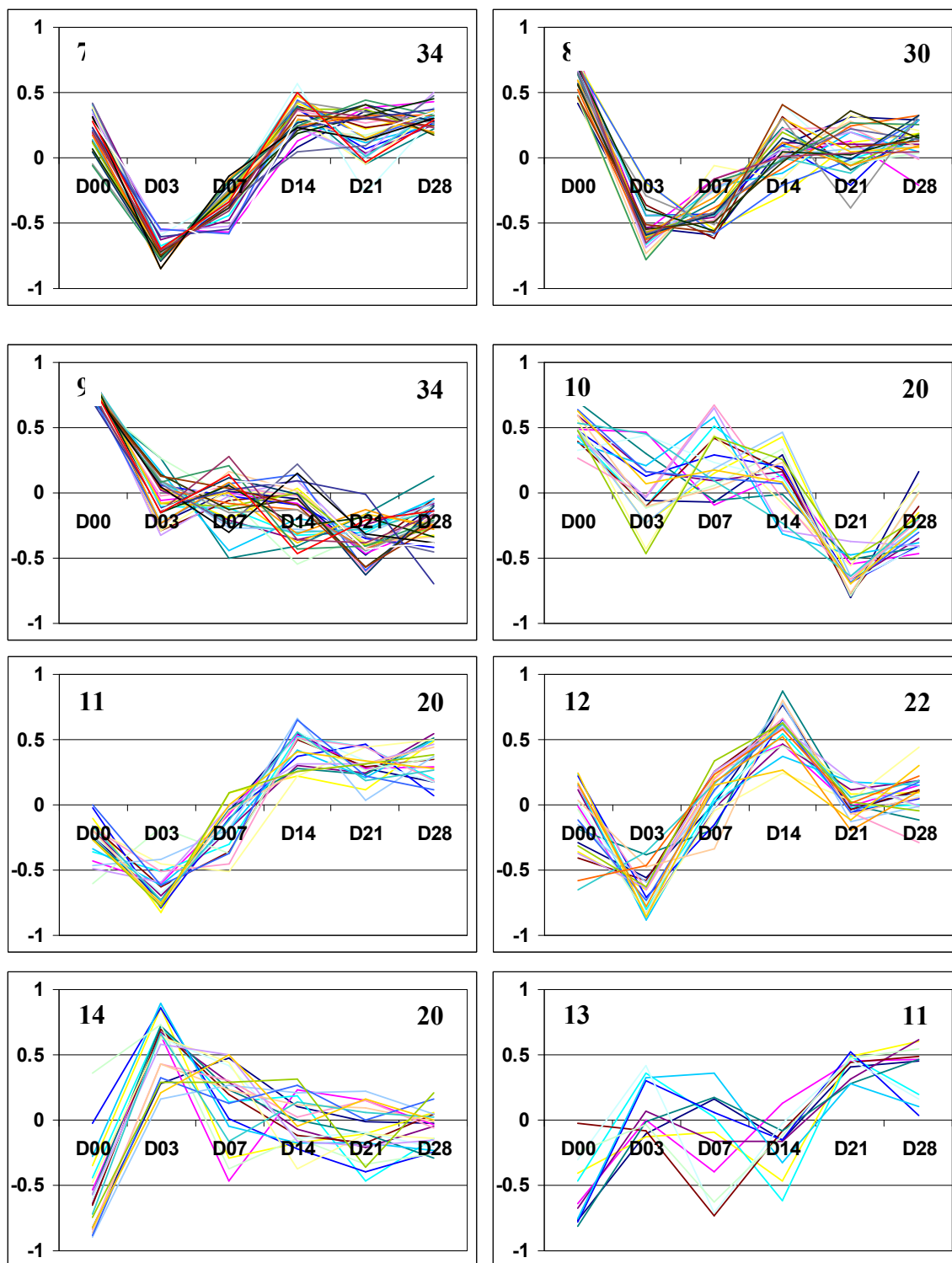


Figure 4.8 Continued

Table 4.2 Real-time PCR measurements of gene expression for select genes

Real-time PCR measurements of expression for select genes at time points 3, 7, 14, 21, 28 referenced against time point 0. All measurements were normalized against dynactin 2.

Gene Description	Time Point				
	3	7	14	21	28
Sema-3A	1.22	-0.42	1.225	-0.615	0.785
Survivin/XIAP	0.045	-1.725	-0.765	-2.905	-1.465
MMP9	1.99	-0.49	-1.855	-1.58	-1.57
IDb1	-0.59	-1.86	-1.635	-0.63	-1.83
EGR1	-4.905	-4.475	-5.18	-6.02	-4.82
PECAM	-2.995	-5.385	-5.78	-4.61	-4.55
SHH	3.015	-0.49	-2.71	-1.11	-2.17
Thromb	1.635	1.855	3.28	4.14	3.25
TGF	0.51	-1.875	-3.365	1.095	-0.58
PDGFB	-2.63	-4.115	-3.835	-1.275	-2.63
Bmp	-2.795	-3.555	-5.695	-0.3	-2.935
SetMynd	-0.167	-0.933	2.5	-0.143	3.63

Principal component analysis (PCA) was also performed in an effort to identify a core set of genetic “programs” within the dataset. PCA is a statistical technique for determining the fundamental variables within a microarray dataset that explain the differences in gene expression between the conditions (timepoints in this case)(Raychaudhuri *et al.*, 2000). PCA was performed using an NIA array analysis tool(Sharov *et al.*, 2005), based on the same mean centered unit normalized data used for agglomerative hierarchical clustering. This analysis demonstrated that 5 principal components (Table 4.3, Figure 4.9) could account for 99.998% of the variability contained within the dataset. Based on the commonly applied criterion of discarding components accounting for less than $(70/n)\%$ of the overall variability, where n is the number of conditions ($70/6 = 11.67\%$), components 4 and 5 would be discarded, however component 4 at 11.124% is very close to the cutoff.

Table 4.3 Principal components of mouse microvessel gene expression data

Results of principal component analysis on the 444 genes identified as differentially expressed. The values for each column are the coefficients of the principal component in relation to each time point. The eigenvalue represents the variance of the component over all genes. The % variance is the percentage of the total variance for which each component accounts.

Projection on Timepoint	Principal Components				
	1	2	3	4	5
0	0.7303	-0.2739	-0.3940	0.2487	0.0886
3	0.2291	0.3972	0.0721	-0.7859	-0.0147
7	-0.0775	0.6998	0.1165	0.5561	-0.1214
14	0.0135	-0.4188	0.8041	0.0767	0.0729
21	-0.5073	-0.1014	-0.3109	-0.0283	0.6843
28	-0.3881	-0.3029	-0.2878	-0.0673	-0.7097
Eigenvalue	42.1723	35.7097	13.6423	12.3554	7.1900
% Variance	37.969%	32.150%	12.282%	11.124%	6.473%

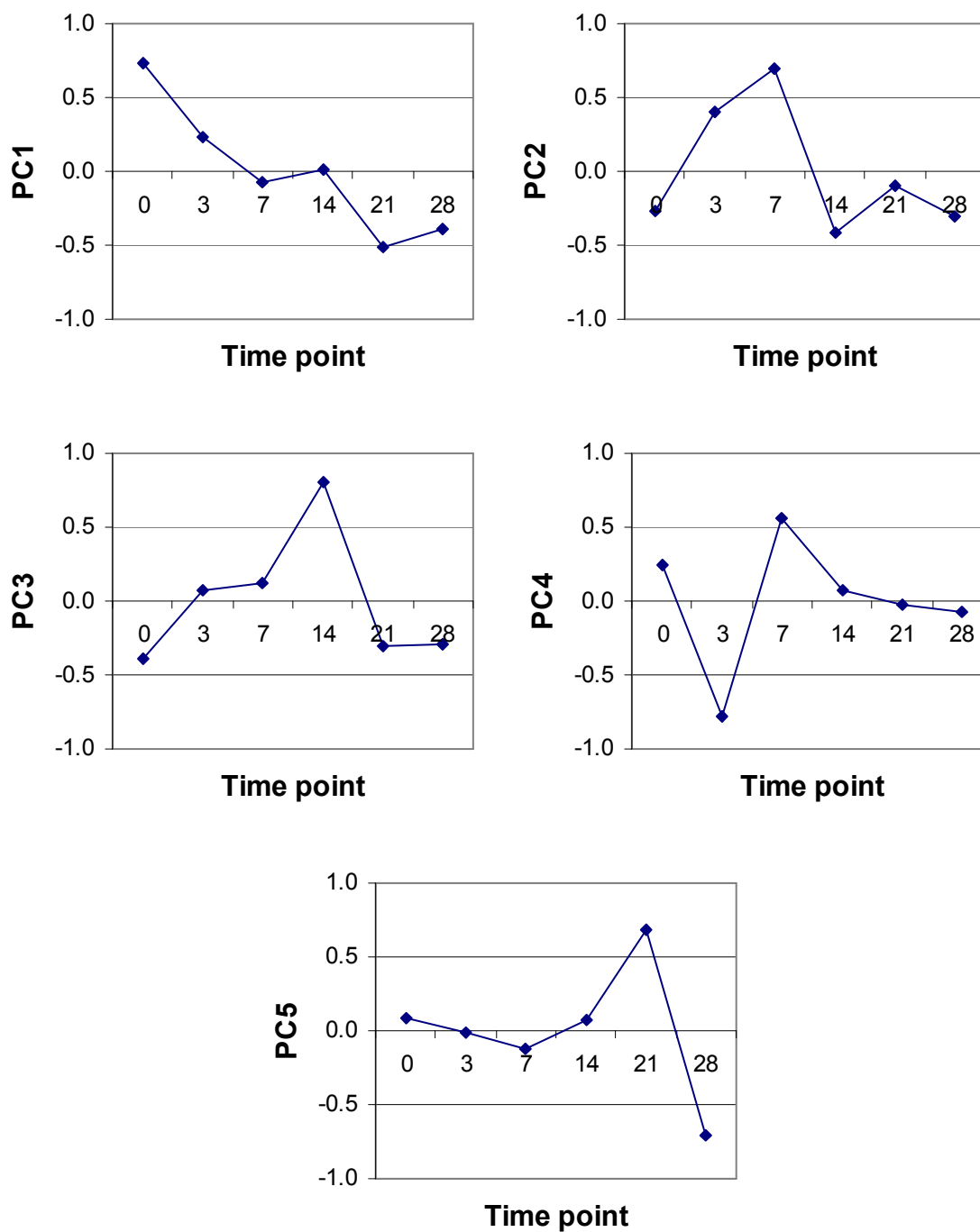


Figure 4.9 Principal components of mouse microvessel gene expression data
Graphical representation of the 5 principal components identified by principal component analysis of the 444 genes identified as differentially expressed.

Another useful technique for interpreting the results of PCA is to map the original observations onto the principal components. This technique allows for the visualization of how all of the time points relate to each other in terms of each principal component. As can be seen in figure 4.10, days 21 and 28 are relatively close to each other in all three components, indicating that the overall pattern of gene expression is similar for these two days. Days 3 and 7 are also somewhat close together in all three components, however not as close together as days 21 and 28. Day 14 is close to the other days in some components but very different in others. Together these results suggest that days 3 and 7 have similar gene expression profiles, a change of expression occurs at around day 14, which then resolves into a final pattern of expression for days 21 and 28.

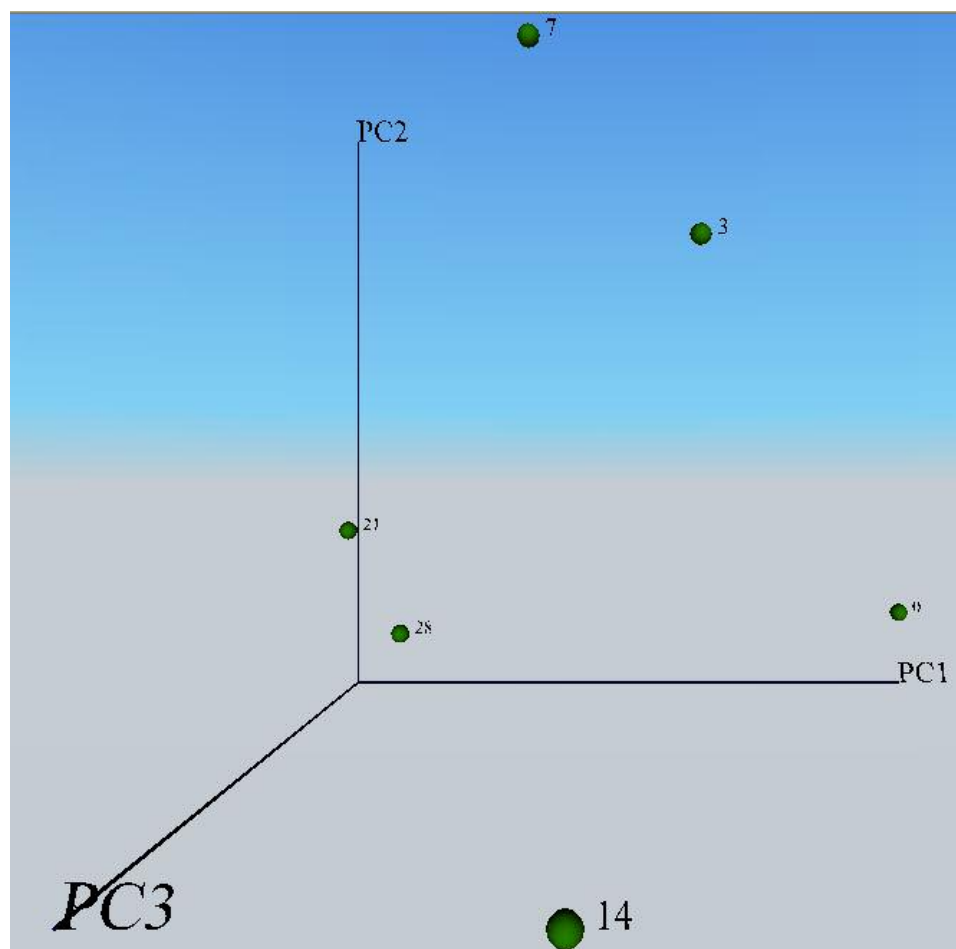


Figure 4.10 Project of the original time points onto first 3 principal components
The projection of the original observations (time points) onto the first 3 principal components.

GoMiner(Zeeberg *et al.*, 2003) was then applied to the microarray dataset in an effort to associate biological information with each gene. This tool looks up the Gene Ontology (GO)(Ashburner *et al.*, 2000) information for each gene, including the biological processes, molecular function, and cellular component with which the gene is associated. GoMiner also assigns a statistical significance to each category based on the percentage of genes in the category that are considered differentially expressed. Table 4.4 contains the list of GO categories that contained a statistically significant increase in the percentage of differentially expressed genes ($p \leq .05$).

Table 4.4 Overrepresented Gene Ontology categories

Gene Ontology categories that were identified as having a statistically significant percentage of differentially expressed genes.

Category Name	P-Chng	Tot	Chng
small GTPase binding	0.0015	11	6
GTPase binding	0.0027	12	6
enzyme binding	0.0029	33	11
cofactor biosynthesis	0.0030	20	8
response to temperature	0.0084	7	4
response to heat	0.0084	7	4
germ cell development	0.0090	4	3
biosynthesis	0.0110	197	38
organelle organization and biogenesis	0.0114	166	33
cytoskeleton organization and biogenesis	0.0125	66	16
cell organization and biogenesis	0.0134	193	37
cytosolic ribosome (sensu Eukaryota)	0.0141	25	8
energy coupled proton transport, down electrochemical gradient	0.0150	8	4
ATP synthesis coupled proton transport	0.0150	8	4
proton-transporting two-sector ATPase complex	0.0150	8	4
cation-transporting ATPase activity	0.0150	8	4
hydrogen-transporting ATP synthase activity, rotational mechanism	0.0150	8	4
hydrogen-transporting ATPase activity, rotational mechanism	0.0150	8	4
phorbol ester receptor activity	0.0185	2	2
protein kinase C activity	0.0185	2	2
dynactin complex	0.0185	2	2
iron ion transport	0.0185	2	2
transmembrane receptor protein tyrosine phosphatase signaling pathway	0.0185	2	2
detection of sound	0.0185	2	2
coenzyme biosynthesis	0.0201	17	6
rhodopsin-like receptor activity	0.0203	5	3
homeostasis	0.0221	22	7
actin binding	0.0226	32	9
cofactor metabolism	0.0226	32	9
cation homeostasis	0.0229	13	5
double-stranded DNA binding	0.0242	9	4
nucleoside phosphate metabolism	0.0242	9	4
ATP biosynthesis	0.0242	9	4
hydrogen transport	0.0242	9	4
nucleoside triphosphate biosynthesis	0.0242	9	4
purine nucleoside triphosphate biosynthesis	0.0242	9	4
ribonucleoside triphosphate biosynthesis	0.0242	9	4
purine ribonucleoside triphosphate biosynthesis	0.0242	9	4
proton transport	0.0242	9	4
purine nucleotide binding	0.0252	227	41
cellular biosynthesis	0.0267	176	33
nucleotide binding	0.0271	228	41
cytoskeleton	0.0317	109	22

oxidative phosphorylation	0.0318	14	5
group transfer coenzyme metabolism	0.0361	10	4
nucleoside triphosphate metabolism	0.0361	10	4
purine nucleoside triphosphate metabolism	0.0361	10	4
purine ribonucleotide biosynthesis	0.0361	10	4
ribonucleoside triphosphate metabolism	0.0361	10	4
purine ribonucleoside triphosphate metabolism	0.0361	10	4
ribonucleotide biosynthesis	0.0361	10	4
di-, tri-valent inorganic cation homeostasis	0.0361	10	4
ATP metabolism	0.0361	10	4
reproductive physiological process	0.0361	10	4
proton-transporting ATP synthase complex (sensu Eukaryota)	0.0365	6	3
lipid catabolism	0.0365	6	3
hydrogen-translocating F-type ATPase complex	0.0365	6	3
proton-transporting ATP synthase complex	0.0365	6	3
transition metal ion homeostasis	0.0365	6	3
mitochondrial membrane	0.0394	35	9
cellular protein metabolism	0.0410	446	72
cell ion homeostasis	0.0425	15	5
macromolecule metabolism	0.0425	503	80
ion homeostasis	0.0425	15	5
cellular macromolecule metabolism	0.0429	482	77
metabolism	0.0447	905	135
protein metabolism	0.0474	449	72
ATPase activity, coupled	0.0474	53	12

4.3 Discussion

Utilizing an in-vivo model of vascularization and a cDNA microarray we have identified 444 genes differentially expressed during the development of a new vascular network. This model recapitulates many of the steps involved in the development of a vascular network, progressing from angiogenesis to vessel maturation and network remodeling. During the initial phase of vascularization, the isolated microvessels sprout and elongate to form a relatively uniform network of small diameter vessels. During this phase proliferation is at its maximum corresponding to a highly angiogenic network. After two weeks some of the vessels appear to have remodeled into larger feed and drain vessels. Over the remaining two weeks this network matures and remodels into a stereotypical vascular bed consisting of larger feed vessels (arteries and arteriols), capillaries, and drainage vessels (veins and venules). In addition, the examination of perivascular cells is also indicative of the initial formation of an immature vascular bed, followed by a progressive maturation into a well differentiated vascular bed. Based on GFP fluorescence measurements, the majority of vessels within the constructs develop from the original isolated microvessels, not the host vasculature. It is also assumed that the associated mural cells are also derived from the original implanted microvessels, as previous work with this model has demonstrated (Shepherd *et al.*, 2004).

Inspection of the list of genes identified as differentially expressed reveals members of some pathways known to be involved in angiogenesis and/or vessel remodeling. For example, platelet derived growth factor (PDGF) and its receptor (PDGFR- β) have been shown to be critical to the recruitment of pericytes to newly

formed vessels(Armulik *et al.*, 2005)(Hoch & Soriano, 2003), and the knockout of either PDGF or PDGFR- β leads to perinatal death due to vascular dysfunction(Leveen *et al.*, 1994)(Soriano, 1994). Consistent with this understanding, the PDGFR- β microarray data displays decreased expression during days 3 and 7, returning to day 0 levels by day 14 and remaining there for the duration of the experiment. Real-time PCR measurements of PDGF also present a similar expression profile. There is also evidence that genes that inhibit angiogenesis play a role in vascular network development. *Brain specific angiogenesis inhibitor* is down regulated during the first week of the time course, and then returns to day 0 levels by day 14, providing further evidence that the model of vascular formation involves an angiogenesis phase that is mostly complete by day 14.

The Gene Ontology (GO) consortium(Ashburner *et al.*, 2000) has developed three structured vocabularies (ontologies) to describe gene products in terms of their biological processes, cellular components, and molecular functions by species. Associations between these ontologies and gene products are designed to represent the existing knowledge about a gene product in each of these three areas. These GO terms are organized using structures called directed acyclic graphs (DAGs) in order to allow lower level terms (more specialized) and gene products to be associated with multiple higher level terms. For example, the molecular function term *protein kinase activity* is associated with both the *protein kinase activity* and *phosphotransferase activity* higher level terms. In order to utilize GoMiner, a list of the gene symbols for all known genes on the microarray and a list of gene symbols for all differentially expressed genes were created. GoMiner was then utilized to look up the GO (biological processes, molecular

function, and cellular component) information for each gene and assign a statistical significance to each category based on the percentage of genes in the category that are differentially expressed. The Gene Ontology categories (Table 4.4) that contained a disproportionate percentage of differentially expressed genes are also consistent with many of the processes known to be involved in the vascularization process. In particular many categories associated with cell proliferation, including metabolism and biosynthesis are present. In addition, the list includes several categories associated with organelle and cytoskeleton organization, consistent with the cell motility and morphological changes necessary for angiogenesis. One category that at first seems inconsistent with our model of vascularization is *germ cell development*. However, this category makes sense given that during the dissection of the epididymal fat it is extremely difficult to exclude all reproductive tissue. In fact, during the isolation process it is often possible to identify tubules that still contain live sperm. Fortunately, all of these cells appear to die off within the first three days (data not shown).

Upon hierarchical clustering, the 444 genes identified as differentially expressed were grouped into 14 clusters. Hierarchical clustering groups genes together based on the similarity of expression values between genes at all timepoints (for a detailed description please refer to chapter 3). Mean centering and unit normalization standardizes and re-centers the gene expression profile for each gene resulting in genes being clustered based on the shape of their expression profile over the timecourse. Therefore, clusters often represent genes that are involved in common or related processes or share some commonly regulatory mechanism. For example, genes involved in cellular metabolism

often change expression levels in response to the metabolic demands of the tissue and exhibit a common expression profile. Also, because RNA was isolated from a tissue containing a heterogeneous population of cell types, it is possible for changes in the measured expression levels of a gene to be due primarily to changes in the proportion of cell types. Unfortunately it is impossible to determine the reason for genes being clustered together based solely on microarray data. Therefore, rather than using the clusters to look for specific relationships between genes, we looked at the overall patterns present in the clusters in the context of our experimental system.

While the hierarchical clustering algorithm generated 14 clusters many of the clusters exhibit similar characteristics. In particular, two common themes were present in many of the clusters. In many of the clusters day 3 is considerably different from day 0. There are a number of possible explanations for this difference including: artifacts due to microvessel isolation and/or suspension in collagen gel, early apoptosis of non-viable microvessels, the new in-vivo environment, early response genes, induction of stress genes, initiation of angiogenesis, etc. The other common cluster theme is that in many clusters gene expression at day 7 is considerable different than day 14. The reasons behind this pattern are not quite as readily apparent as differences between day 0 and 3, however in combination with the perfusion data, we believe that initiation of blood flow returns angiogenesis related genes to their quiescent state and initiates changes in the transcription of maturation and remodeling genes. Hierarchical clustering also demonstrated that most genes either had one peak or trough in expression level, or continuously increased or decreased in expression after day 3, indicating that most of the

differentially expressed exhibit alternate levels of expression during the angiogenesis and maturation phases of vascularization. In other words, very few genes were either consistently up or down regulated during the entire vascularization process.

Principal component analysis identifies overall trends in the dataset (Figure 4.9). Component 1 represents the average gene expression at each time point. Its overall downward trend suggests that as time progresses, gene expression becomes more divergent. It is important to remember that each gene can have a positive or negative value for each principal component, therefore component 1 does not mean that on average gene expression goes down over time, rather it means that on average gene expression tends to either go up and continue going up or down and continue going down, over time. Component 2 represents the dominant trend in the data after component 1 has been taken into account. This component shows that once the overall up or down trend has been removed, gene expression tends to either be higher or lower at both days 3 and 7, as compared to the remaining days. We interpret this component to be indicative of the angiogenic phase of the development of the network of blood vessels. Component 3 identifies that once components 1 and 2 have been taken into account, gene expression tends to be most divergent at day 14. We believe that this difference in gene expression is due to the onset of flow, and represents a transition from angiogenesis to blood vessel maturation and remodeling. Components 4 and 5 would typically not be considered because they explain less than $(70/n)$ percent of the total variability, however they are indicative of differential expression specific to days 3, and 21 respectively. In particular, component 4, which is very close to the cutoff for significance, indicates that once all of

the other components have been removed, there is still considerable unexplained variation between day 3 and the other days. Evidence for the significance of this component is provided by the similar pattern present in many of the hierarchical clusters.

Taken together, these molecular and morphometric characterizations of the implanted microvascular constructs point to a highly regulated process of vascular network development in which an initial phase of angiogenesis is followed by vessel maturation and network remodeling. This switch from network expansion to network adaptation appears to correspond with the initiation of blood flow and a change in the overall pattern of gene expression. Before the onset of blood flow, a network of small diameter vessels develops that is consistent with microvessel networks that develop in the in-vitro environment(Hoying *et al.*, 1996). This observation begs the question of whether it is the in-vivo environment or the initiation of flow (or both) that direct the maturation and remodeling of the vascular network. Unfortunately, it is impossible to segregate those two conditions using this in-vivo experimental model. This model does however indicate that there are separate programs of gene expression associated with the angiogenesis phase and maturation phase of vascularization. In addition, given the significant number of genes that display changes in gene expression at day 14, and the identification of a principal component specific to day 14, it appears that there is a unique program associated with the onset of flow.

5 CONCLUSION

High-throughput techniques such as microarrays are now making it possible to monitor experimental systems at an unprecedented level of resolution. In fact, it is now feasible to measure the expression level of the majority of genes in any given sample for a variety of different organisms. This dramatic increase in scale from a few genes to tens of thousands of genes has demanded a change in the approaches to both experimental design and data analysis. Performing thousands of simultaneous measurements presents numerous challenges including the impracticality of monitoring each measurement, suboptimal experimental conditions for each probe, and the high probability of obtaining one or more erroneous measures. However, this large number of measurements confers some advantages as well. In particular, microarray experiments usually suffer from a relatively small number of measurements for each gene, however thousands of genes are measured under almost identical environmental conditions (i.e. in the same hybridization). Therefore, it is possible to incorporate some of the information from all of the genes on the microarray into the analysis of each gene on the array.

One of the most vital, but often underemphasized, steps in any experiment is the design of the experiment. This process is particularly important in multi-channel microarray experiments, where both the sample population and hybridization scheme can dramatically affect the conclusions that can be drawn from the resulting data (Yang & Speed, 2002). When selecting a procedure for obtaining samples it is imperative to consider the population from which the experimental samples will be derived. This sample population not only defines the population to which a result may be applicable,

but can dramatically impact the ability to detect statistically significant differences. For example, performing an experiment using one cell type, by definition limits the interpretation of the results to just that one type of cell. In fact, in many cases the interpretation of an experiment may actually be limited to a particular passage, or even a specific aliquot of cells. However, performing an experiment using a more diverse sample population (e.g. different types of cells, or cells from different organisms) may make it impossible to identify statistically significant experimental differences. In particular, all experiments, and especially microarray experiments should incorporate some form of biological replication. Biological replication refers to samples prepared from separate sources – for example from individual mice, or separate flasks of cells. This is opposed to technical replication, where the same samples are used in multiple hybridizations. Practically speaking it is necessary to consider the cost, time, and availability of samples and microarrays, as well as both experimental and biological variability, when designing a robust microarray experiment.

It is imperative to define, and consider, the statistical analysis that will be performed on the resulting dataset. In general, for a fixed number of hybridizations, designs with more biological replicates per treatment/condition and fewer hybridizations per replicate are more powerful than designs with fewer biological replicates and more hybridizations per replicate (Cui & Churchill, 2002). In addition, designs such as the loop design, which incorporate direct comparison, are more efficient than reference designs because no measurements are wasted on the reference sample (Qin & Kerr, 2004). When comparing two samples, the technical variance of a direct comparison design is one-

fourth of that obtained using a reference design based on the same number of hybridizations(Churchill, 2002).

Before performing a statistical analysis of microarray data it is first necessary to transform and pre-process raw microarray data. Microarray image processing software generates a multitude of measurements for each spot from which one measure of intensity is calculated. The most controversial aspect of this process is whether to subtract some measure of background intensity from the intensity of each spot. Background subtraction has the advantages of reducing special bias and improving estimates of gene expression by subtracting fluorescent signal that is unrelated to the fluorescently labeled hybridized target cDNA. Some researchers have advocated against background subtraction due to its tendency to increase variability for low intensity measurements. This increase in variability, however, is not caused by the background subtraction, but is a result of performing a simple log transformation on the background subtracted intensities. A log-based transformation is often used because it confers two advantages; it allows the use of an additive linear model, and it accounts for the multiplicative error that is present over the majority of the range of measured values. A simple log transformation, however, is inappropriate because it greatly increases variability for small values, where additive error dominates due to the inability of microarray scanners to accurately distinguish small differences in intensity. Implementing a mixed linlog transformation that performs a log transformation for large numbers, and a linear transformation for small numbers maintains the advantages of a log-based transformation and incorporates an appropriate transformation for small numbers.

Because of the significant variability between hybridizations, microarray data must also be normalized. Normalization between the two channels of each hybridization accounts for differences in reverse transcription efficiencies, fluorochrome characteristics, scanner sensitivity and settings, etc. Initial transformations were based simply on average differences between each channel (wavelength), however both spatial and intensity biases have been observed by many researchers (Yang *et al.*, 2002b; Yang *et al.*, 2002a; Cui *et al.*, 2003; Smyth & Speed, 2003; Wilson *et al.*, 2003). To account for these possible biases we incorporated a locally weighted regression (lowess), which includes terms for both the intensity and the location of the spot on the array, into our analysis.

Microarrays are used to quantitate differences in the levels of mRNA transcripts (or genomic DNA) between two or more samples, however these differences are derived from fluorescence based measurements. Each step in the preparation and hybridization of the samples, as well as the fluorescence imaging, and data extraction process can introduce variability and/or bias into the measured intensities. Researchers should make every effort to minimize variability by performing all sample preparation and hybridization in as short of period of time as possible, using the same reagents, by the same technician. However, even if all possible precautions are taken, and the data is preprocessed to remove systematic biases, there will still be considerable variability and bias in any microarray dataset. Fortunately, employing a statistical model incorporating known sources of variability significantly improves the ability to detect differentially expressed genes. In particular, a linear model incorporating terms to account for array,

dye (fluorochrome), and gene variability greatly improves the ability to identify genes differentially expressed between treatments/conditions. Employing an analysis of variance (ANOVA) in combination with an additive linear model provides better estimates of differences in gene expression as well as provides a measure of significance for those differences.

In addition to simply identifying differentially expressed genes, patterns of expression within large scale gene expression datasets can be used to identify relationships between genes. Agglomerative hierarchical clustering was the first, and still most used, clustering algorithm to be applied to microarray data. Agglomerative hierarchical clustering is actually a class of clustering algorithms that each can employ one or more distance metrics. The results of these algorithms are intuitive and no advanced knowledge of the number of clusters in the dataset is required, however the investigator must select the clustering algorithm and distance metric to be employed. In an effort to evaluate the performance of we developed software to generate simulated microarray datasets containing known clusters, perform the clustering, and evaluate the ability of each cluster/distance metric combination to recover the assigned clusters. In total nine different clustering algorithms were assessed, included the three most commonly used algorithms (nearest neighbor, furthest neighbor, and average linkage), using four different distance metrics.

The performance of the nine clustering algorithms in recovering the known clusters within the simulated data varied dramatically. The *nearest neighbor* algorithm was by far the worst performer due to its tendency to inappropriately chain observation

together. At the other end of the spectrum, the *incremental sum of squares* and *flexible β* algorithms performed consistently well irrespective of the distance metric employed. In general the distance metric utilized had less of an impact on clustering performance than the choice of clustering algorithm. The distance metric can however, dramatically affect clustering results if mean centering and unit normalization is not performed first.

Specifically, *Pierson's correlation coefficient* disregards the offset and magnitude of each gene's expression profile, clustering genes based on the shape of their expression profile. Therefore if one wants to take into account either magnitude or offset during clustering *Pierson's correlation coefficient* should not be used. The three most commonly used algorithms for clustering gene expression profiles are still hierarchical clustering, k-means clustering, self-organizing maps. However both these techniques, and more recent advanced techniques that attempt to capture the coherence of a subset of genes or conditions (Cheng & Church, 2000) (Tanay *et al.*, 2002) (Sheng *et al.*, 2003) (Kluger *et al.*, 2003), only consider the mathematical or statistical similarity of gene expression profiles.

The term angiogenesis is often used to describe the process of forming new blood vessels from existing blood vessels, however this single term does not adequately describe the process of new blood vessel formation, let alone the formation of a vascular network (Pepper *et al.*, 1992). Angiogenesis occurs under normal conditions during embryonic development, the female reproductive cycle, and wound healing, as well as a component of pathological conditions such as solid tumor cancers, hemangiomas, arthritis, psoriasis, and pyogenic granulomas (Zetter, 1988). However, even a cursory comparison between the vessels that are formed during these varied processes, especially

as part of many pathological conditions, demonstrates that all angiogenesis-derived vessels are neither physiologically nor functionally equivalent. In particular, the vasculature associated with most solid tumor cancers is both structurally and functionally abnormal (Jain, 2005). Recent research also indicates that tumor vasculature is stuck in an angiogenic state and that the synergistic effect of chemotherapy and antiangiogenic drugs is due to the normalization of the tumor vasculature, rather than the destruction of the vasculature (Winkler *et al.*, 2004). Both the morphological and molecular data presented in this dissertation support a model of angiogenesis-based vascular development that begins with an angiogenesis phase, wherein an initial tube is formed that is capable of carrying blood, followed by a maturation and remodeling phase. During this second phase the newly formed vessels structurally adapt (including vessel regression) to meet the metabolic demands of the tissue and accommodate the necessary flow of blood.

The development of a vascular network is a complex process involving multiple cell types, which must proliferate, communicate, and assemble into a highly organized structure capable of adapting to changes in the perfusion requirements of tissue and changes in blood pressure and flow. Significant previous work in this field, focusing mainly on angiogenesis, has uncovered some of the molecular mechanisms regulating vascularization, however utilizing classic (pre-genomic) techniques it is only possible to evaluate the expression levels of a few genes at a time. In an effort to understand some of the genetic mechanisms responsible for regulating the development of a vascular

network, we manufactured a mouse cDNA array containing the NIA 15,000 mouse cDNA cloneset for use with our mouse microvessel model of vascularization.

An obvious, but often overlooked, attribute of microarray technology is that all cells within the samples contribute to the measured levels of transcription. This detail makes it very difficult to assess transcription during vascularization because vascular cells usually represent only a small fraction of the cells in any tissue. Utilizing a mouse microvessel model of vascularization, in which small vessels are isolated from adipose tissue, resuspended in collagen, and then implanted into the hindquarters of SCID mice we were able to obtain samples, which consisted of mostly vascular cells, for the different phases of the vascularization process. These samples explanted after 3, 7, 14, 21, and 28 days of implantation exhibited morphological and molecular characteristics consistent with a vascular development process that begins with a highly angiogenic phase, which arrests when blood flow begins, and is followed by a maturation and remodeling phase.

RNA from samples that were explanted after 3, 7, 14, 21 and 28 days, and freshly isolated vessels, was comparatively hybridized using the ~15,000 clone mouse cDNA microarray. 3,470 genes were expressed at high enough levels to be measured confidently for at least one time point, and 444 of these genes were identified as differentially expressed between time points. Approximately half of the differentially expressed genes were identifiable based on sequences contained in Genbank. This limited number of genes made it difficult to address our original goal of elucidating the genetic mechanisms that coordinate the vascularization process, however evaluation of

overall patterns of gene expression implicate different genetic programs for the angiogenesis and maturation phases.

Hierarchical clustering segregated the differentially expressed genes into 14 clusters, however many of these clusters exhibited similar profiles. Principal component analysis then revealed that 3 components could account for 82% of the variability contained within the microarray dataset. After accounting for the overall average levels of expression at each time point (component 1), components 2 and 3 exhibited patterns consistent with an angiogenic phase and maturation/remodeling phase respectively. Component 3 also indicates that there is a component of gene expression that is unique to day 14, which corresponds with the initiation of blood flow.

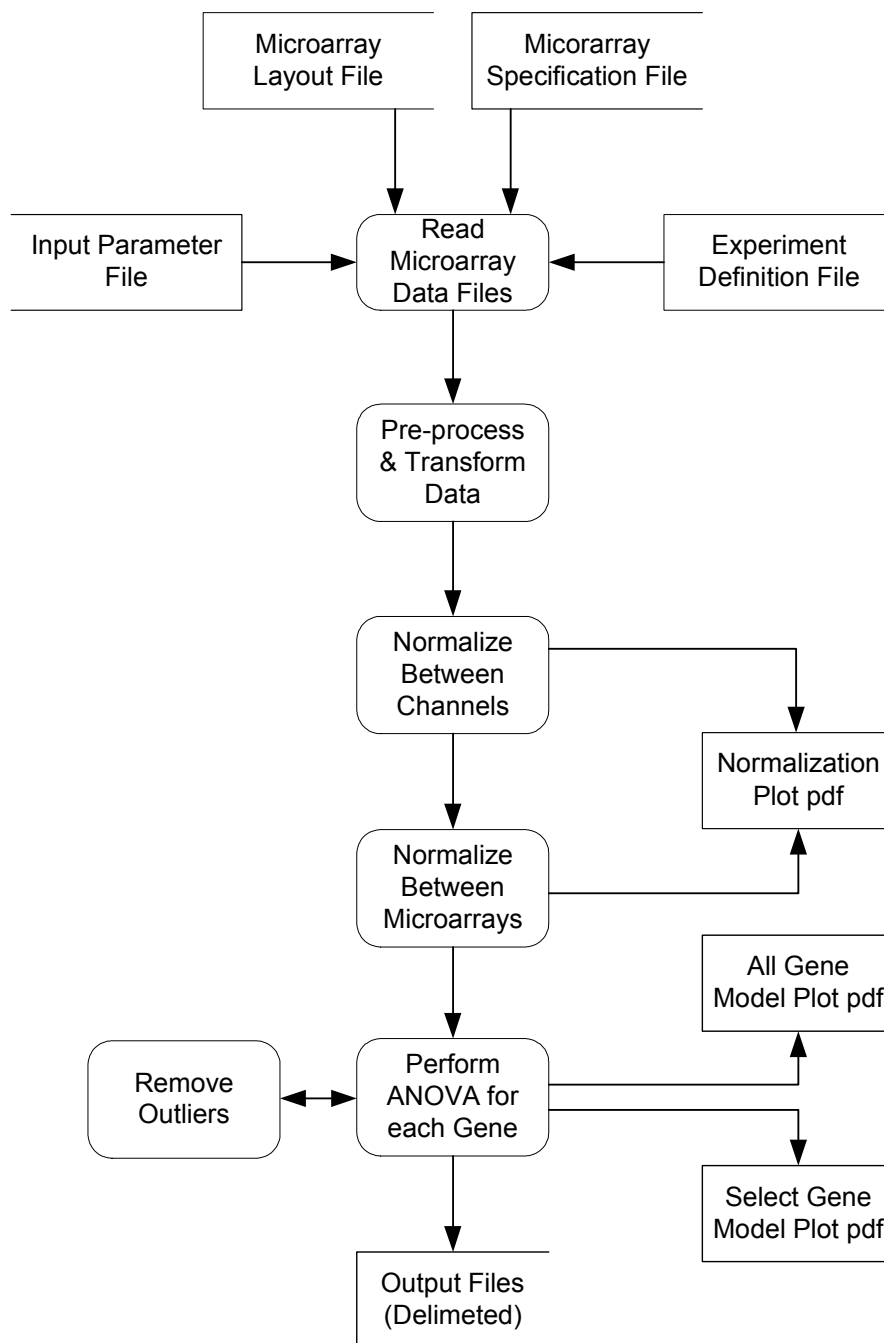
Genomics and high throughput techniques such as microarrays have dramatically impacted the way most scientific research is conducted. They are however, not a substitute for good experimental design and robust data analysis. Microarray data in particular is prone to over interpretation as it is always possible to identify subsets of data that support one's hypothesis. In addition, current microarray technology does not provide sufficient sensitivity to measure the expression levels of low abundance transcripts. Microarray are especially sensitive to small differences between samples, due to the number of genes being measured, and therefore stringent experimental control and biological replication are essential.

Currently there are four primary uses for microarrays: gene identification, global pattern recognition, pathway reconstruction, and sample classification. Microarrays can be used successfully for gene identification as long as it is understood that only relatively

highly expressed genes will be identified, and each finding should be confirmed using an alternate method such as real-time PCR. Global pattern recognition has been used successfully in cases where there are dramatic changes in cellular processes, such as sporulation. Pathway analysis is the desired outcome of many microarray experiments but limitations in microarray technology and difficulties in data interpretation have limited the success of many of these types of experiments. Sample classification has proven to be the most applicable use for microarray technology because it does not require the measurement of specific genes, and only a subset of the genes involved in any given process are required to discriminate between samples. In particular, microarrays have demonstrated the ability to accurately distinguish between multiple forms of cancer(Alizadeh *et al.*, 2000)(Ramaswamy *et al.*, 2001).

APPENDICES

APPENDIX A DATA FLOW DIAGRAM FOR CARMA



CARMA utilizes four input files to direct each analysis. The *microarray layout* and *microarray specification* files are common to all experiments that use a specific

microarray and describe the layout of the spots on the array and the details about each spot, respectively. The *input parameter* and *experiment definition* files are specific to each experiment and contain the parameters to be used in the analysis and the sample hybridization scheme, respectively. Each analysis begins with the reading of these four input files by CARMA. Each hybridization is then processed individually to perform the linlog transformation of the measured intensity values and flag spots with insufficient intensities (as specified in the *input parameter* file). Normalization is then performed between the two channels of each array, followed by normalization between both channels of each array and the average of both channels of all arrays in the experiment. Both a numeric and graphical representation of each of these processes for each hybridization is written to output files. Next a gene-by-gene analysis of variance (ANOVA) is performed for all genes that a measured confidently for a minimum number of hybridizations for at least one sample (also specified in the *input parameter* file). During the ANOVA outliers are removed (if so desired) based on the inconsistencies within the replicate measures. Outliers are removed recursively until the significance of the most extreme outlier is above the value specified in the *input parameter* file. The results of the ANOVA for each gene analyzed are written to output files in both numeric and graphical formats.

**APPENDIX B AGGLOMERATIVE HIERARCHICAL CLUSTERS FOR MOUSE
MICROVESSEL EXPERIMENT**

Cluster 1 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
H3052A08	0.00	1.80	1.05	2.12	1.74	0.56	0.54	0.57	0.56	0.54
H3057A02	-0.41	0.55	0.25	1.48	1.04	0.36	0.34	0.36	0.36	0.36
chromodomain helicase DNA binding protein 1-like	-0.06	0.65	0.33	1.26	1.17	0.34	0.33	0.35	0.34	0.33
H3061A09	-0.42	0.31	0.45	0.85	1.11	0.24	0.24	0.25	0.24	0.23
H3062B03	-0.20	0.31	0.59	1.04	1.03	0.29	0.29	0.30	0.29	0.28
H3013A10	-0.02	0.22	0.47	1.08	0.88	0.24	0.23	0.25	0.24	0.24
RIKEN cDNA 4930404N11 gene	0.07	1.23	1.24	1.45	1.46	0.43	0.42	0.43	0.43	0.41
Laminin B1 subunit 1	-0.07	1.30	0.87	1.90	1.50	0.38	0.37	0.39	0.38	0.37
H3033A06	-0.19	0.88	0.68	1.65	1.28	0.45	0.45	0.45	0.44	0.42
H3049A06	-0.36	0.64	1.04	1.47	1.15	0.35	0.34	0.35	0.35	0.34
H3035B12	0.21	1.34	1.08	1.63	1.71	0.46	0.45	0.47	0.46	0.45
H3040B12	-0.14	0.61	0.28	1.36	1.12	0.34	0.33	0.35	0.34	0.33
ganglioside-induced differentiation-associated- protein 1	0.46	1.11	1.11	1.96	1.21	0.40	0.40	0.41	0.40	0.40
expressed sequence AA409316	-0.07	1.69	1.01	1.74	1.41	0.54	0.52	0.54	0.54	0.52
expressed sequence A1642036	-1.04	0.35	0.14	1.01	0.85	0.39	0.35	0.37	0.36	0.35
RIKEN cDNA 1500004A08 gene	-0.56	0.40	0.73	1.12	0.64	0.36	0.33	0.34	0.33	0.32
H3003D09	-0.13	0.75	0.76	1.27	1.27	0.33	0.33	0.34	0.33	0.32
H3148D03	0.14	1.63	0.77	1.93	1.76	0.57	0.56	0.58	0.57	0.55
H3058C10	-0.32	0.56	0.60	1.00	1.19	0.35	0.34	0.35	0.35	0.34
H3106C04	0.24	1.40	0.92	1.74	1.31	0.35	0.33	0.34	0.33	0.32
RIKEN cDNA 1700001L05 gene	0.08	0.88	0.31	1.40	1.08	0.33	0.33	0.33	0.33	0.32
H3008C05	0.33	1.65	1.10	1.97	1.66	0.34	0.33	0.34	0.34	0.32
RIKEN cDNA A530058O07 gene	0.01	0.90	0.62	1.81	1.57	0.47	0.45	0.47	0.47	0.45
RAB3D, member RAS oncogene family	-0.07	1.33	0.95	1.67	1.29	0.43	0.42	0.44	0.43	0.42
actin-binding LIM protein 1	0.24	1.03	0.79	1.34	1.54	0.33	0.33	0.33	0.33	0.32
H3033C12	-0.67	0.20	0.29	1.75	1.59	0.44	0.43	0.46	0.45	0.43
RIKEN cDNA D230025D16 gene	-1.32	0.75	0.35	1.02	0.74	0.34	0.30	0.32	0.31	0.30
dynactin 5	0.36	1.06	0.59	1.73	1.34	0.41	0.40	0.41	0.41	0.39
H3035D06	-0.17	0.82	0.57	1.58	1.33	0.42	0.41	0.42	0.42	0.40
H3047D12	0.31	1.19	0.87	1.87	1.34	0.45	0.44	0.46	0.45	0.44
polymerase (DNA directed), epsilon 2 (p59 subunit)	0.60	0.90	0.44	1.71	1.24	0.32	0.31	0.33	0.32	0.31

RIKEN cDNA 1500016L11 gene	0.28	1.18	0.79	1.71	0.71	0.30	0.30	0.30	0.30	0.30
H3047F10	-0.28	0.79	0.58	1.54	1.07	0.40	0.39	0.41	0.40	0.39
protein phosphatase 4, regulatory subunit 2	0.27	1.00	0.45	1.24	0.94	0.29	0.28	0.30	0.29	0.28
RIKEN cDNA 6820402O20 gene	0.28	1.54	0.72	1.86	1.62	0.48	0.46	0.48	0.48	0.46
Z3001F10	-0.57	1.05	0.90	1.78	1.38	0.33	0.32	0.34	0.33	0.32
RIKEN cDNA C330039G02 gene	-0.43	0.62	0.95	1.60	1.25	0.36	0.36	0.36	0.36	0.35
H3034F11	-0.05	0.70	0.62	1.44	0.99	0.34	0.33	0.34	0.34	0.33
H3038F05	0.24	1.24	0.69	1.78	1.29	0.34	0.33	0.34	0.34	0.32
H3048F05	-0.30	1.15	0.92	1.43	1.52	0.51	0.50	0.52	0.51	0.49
coenzyme Q4 homolog (yeast)	-0.09	1.61	1.10	1.94	1.50	0.45	0.46	0.46	0.45	0.44
H3037E12	0.05	1.18	0.84	1.77	1.31	0.39	0.38	0.40	0.39	0.38
phospholipase C, delta 1	0.08	0.87	0.64	1.34	1.20	0.27	0.26	0.27	0.27	0.26
malignant fibrous histiocytoma amplified sequence 1	-0.08	1.20	0.91	2.09	1.72	0.30	0.29	0.31	0.30	0.29
H3037F12	-0.42	0.48	0.65	1.10	0.63	0.22	0.22	0.23	0.22	0.21
H3005H01	0.04	0.97	0.78	1.25	0.87	0.29	0.28	0.30	0.29	0.29
H3039G08	0.10	1.23	1.01	1.54	1.20	0.41	0.39	0.41	0.41	0.39
RIO kinase 2 (yeast)	-0.29	0.95	0.77	1.91	1.45	0.47	0.46	0.48	0.47	0.45
H3035H02	0.15	0.98	1.26	1.90	1.53	0.47	0.47	0.48	0.48	0.46
H3038G09	-0.10	1.35	1.14	1.77	1.65	0.44	0.42	0.44	0.44	0.42
basic leucine zipper and W2 domains 2	0.45	1.08	1.11	1.94	1.88	0.28	0.27	0.29	0.28	0.27
RIKEN full-length enriched library, clone:B930009B01	-0.90	-0.20	-0.33	0.23	1.03	0.38	0.36	0.38	0.37	0.37
DNA segment, Chr 8, ERATO Doi 594, expressed	-0.22	0.78	0.69	1.38	1.09	0.25	0.24	0.26	0.26	0.24
X-ray repair complementing defective repair in Chinese hamster cells 5	-0.37	0.93	0.53	2.06	1.38	0.42	0.41	0.43	0.42	0.41
H3087G11	0.12	1.45	1.08	1.95	1.52	0.32	0.31	0.32	0.31	0.30
H3038G12	-1.10	0.30	0.11	1.04	0.68	0.44	0.42	0.44	0.44	0.42

Cluster 2 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
RIKEN cDNA E130008O17 gene	-0.58	1.06	0.83	1.38	1.32	0.53	0.51	0.53	0.53	0.51
H3055B01	-0.69	1.26	0.33	1.12	0.69	0.47	0.46	0.48	0.47	0.46
H3084B01	-0.85	1.37	1.00	1.52	0.94	0.59	0.58	0.60	0.59	0.58
RIKEN cDNA 3110003A17 gene	-0.76	0.98	0.55	0.82	0.76	0.42	0.40	0.42	0.42	0.40
H3089A08	-0.47	1.18	0.66	1.34	0.84	0.35	0.34	0.36	0.35	0.35
H3086B02	-0.52	0.94	0.52	0.88	0.47	0.30	0.30	0.31	0.30	0.30
RIKEN cDNA 2610027L16 gene	-0.13	1.45	0.82	0.75	0.50	0.27	0.26	0.27	0.27	0.26
H3047A03	-0.52	1.19	0.92	0.67	1.01	0.44	0.42	0.44	0.44	0.42
ribosomal protein L9	-0.80	1.23	0.77	1.65	1.23	0.46	0.45	0.47	0.46	0.45
CUG triplet repeat, RNA binding protein 1	-0.43	0.65	0.15	1.08	0.52	0.29	0.28	0.29	0.28	0.27
H3097A10	-0.11	1.35	0.46	1.24	0.94	0.40	0.38	0.40	0.40	0.38
H3050B04	-1.00	1.18	0.68	0.35	0.31	0.48	0.47	0.49	0.48	0.47
H3109B04	-0.41	0.95	0.62	1.27	0.95	0.35	0.34	0.35	0.35	0.34
translocase of inner mitochondrial membrane 9 homolog (yeast)	-0.50	0.98	0.49	1.50	0.77	0.42	0.41	0.43	0.44	0.41
ATPase, H ⁺ transporting, V0 subunit C	0.48	1.59	1.50	1.46	1.66	0.41	0.39	0.41	0.41	0.39
solute carrier family 35, member A5	-0.86	1.56	0.92	1.79	1.18	0.44	0.44	0.45	0.44	0.44
ring finger protein 24	-0.32	1.34	0.84	1.77	1.37	0.43	0.42	0.44	0.43	0.42
peroxiredoxin 1	-0.32	1.02	0.62	1.36	0.82	0.38	0.37	0.39	0.38	0.37
checkpoint with forkhead and ring finger domains	0.00	1.62	0.87	1.22	1.33	0.43	0.42	0.43	0.43	0.42
cDNA sequence BC035044	1.31	3.38	3.67	2.97	3.84	0.84	0.81	0.85	0.84	0.81
c-myc binding protein	-1.26	0.79	0.86	0.55	0.64	0.52	0.50	0.53	0.52	0.50
RAS-related C3 botulinum substrate 1	-1.16	1.07	1.08	0.42	0.79	0.47	0.45	0.47	0.47	0.45
claspin homolog (Xenopus laevis)	0.15	1.82	1.15	1.48	1.53	0.46	0.45	0.47	0.46	0.45
H3018C10	-0.79	1.07	0.75	0.11	0.49	0.36	0.35	0.37	0.36	0.35
RIKEN cDNA 2210409B22 gene	-0.35	1.46	0.99	1.93	1.51	0.41	0.40	0.42	0.41	0.40
RIKEN cDNA D230025D16 gene	-0.43	1.23	0.72	1.27	1.11	0.43	0.42	0.43	0.43	0.41
2,3-bisphosphoglycerate mutase	-0.22	1.54	1.17	1.77	1.52	0.46	0.45	0.47	0.46	0.45
ATP synthase, H ⁺ transporting, mitochondrial F0 complex, subunit b, isoform 1	-0.44	0.99	0.53	0.82	0.76	0.33	0.32	0.33	0.33	0.32

Z3003E11	-0.56	1.18	0.79	1.47	1.22	0.44	0.42	0.44	0.44	0.42
H3037F05	-0.20	1.20	0.67	1.56	1.21	0.43	0.42	0.44	0.43	0.42
eukaryotic translation initiation factor 3, subunit 10 (theta)	0.82	1.49	1.46	1.80	1.51	0.24	0.24	0.25	0.24	0.23
heat shock protein 1, alpha	-0.81	0.89	0.75	0.60	0.43	0.21	0.21	0.21	0.21	0.20
H3047H02	-0.53	1.36	0.84	1.44	1.07	0.38	0.37	0.39	0.38	0.37
H3003G03	-0.35	1.18	0.90	1.77	1.28	0.36	0.35	0.36	0.36	0.35
brain abundant, membrane attached signal protein 1	-0.60	1.25	0.92	1.16	1.10	0.38	0.36	0.38	0.38	0.36
H3045G03	-0.74	1.26	0.96	1.01	0.69	0.47	0.46	0.48	0.47	0.46
ribosomal protein L37a	-1.36	0.95	0.63	-0.12	0.06	0.54	0.53	0.55	0.54	0.53
Ras suppressor protein 1	-0.36	1.28	0.83	0.97	0.57	0.33	0.32	0.33	0.33	0.31
H3045G04	-0.68	1.58	1.68	0.01	0.77	0.72	0.70	0.73	0.72	0.70
H3047G10	-0.18	1.39	0.95	1.21	0.98	0.39	0.38	0.40	0.39	0.38
H3048G10	0.12	1.45	0.88	1.18	1.14	0.33	0.32	0.33	0.33	0.32
H3111G04	0.51	1.41	1.03	1.13	1.04	0.27	0.26	0.27	0.27	0.26
H3016H04	-0.37	0.77	0.67	0.72	0.73	0.22	0.21	0.23	0.22	0.21
H3035H04	-0.68	0.87	0.26	0.83	0.55	0.34	0.33	0.34	0.34	0.33
Down syndrome critical region gene 1-like 2	-0.27	1.07	0.27	0.45	0.95	0.32	0.31	0.33	0.32	0.31

Cluster 3 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
H3037A01	-3.12	-1.21	-2.10	-2.18	-2.66	0.70	0.68	0.71	0.70	0.68
a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 1	-3.52	-2.08	-2.96	-2.76	-2.93	0.69	0.67	0.70	0.69	0.67
tribbles homolog 3 (Drosophila)	-3.54	-1.95	-2.93	-3.14	-3.24	0.64	0.62	0.65	0.64	0.62
DEAD (Asp-Glu-Ala-Asp) box polypeptide 19	-2.48	-0.99	-1.96	-1.73	-1.87	0.47	0.46	0.48	0.47	0.45
sphingosine kinase 2	-3.88	-2.10	-2.83	-3.10	-2.77	0.73	0.71	0.74	0.73	0.71
H3062A09	-2.65	-1.62	-2.42	-2.40	-2.28	0.46	0.45	0.47	0.46	0.45
H3065A09	-2.10	-1.31	-1.86	-1.75	-1.54	0.43	0.42	0.44	0.43	0.42
tyrosine kinase, non-receptor, 1	-3.19	-1.86	-2.69	-2.47	-2.23	0.63	0.61	0.64	0.63	0.61
H3048B09	-3.25	-1.66	-2.62	-2.81	-2.70	0.62	0.60	0.62	0.61	0.59
adaptor protein complex AP-1, gamma 1 subunit	-3.18	-1.43	-2.46	-2.78	-2.82	0.56	0.54	0.57	0.56	0.54
Mus musculus upstream transcription factor 2 (Usf2), mRNA	-3.32	-1.90	-2.53	-2.81	-2.87	0.56	0.55	0.57	0.56	0.55
H3031B11	-3.26	-1.56	-2.41	-2.73	-2.72	0.60	0.58	0.61	0.60	0.58
ATP-binding cassette, sub-family B (MDR/TAP), member 10	-2.99	-1.77	-2.52	-2.58	-2.53	0.54	0.53	0.55	0.54	0.52
Minichromosome maintenance deficient 3 (S. cerevisiae) associated protein	-3.40	-1.87	-2.74	-2.60	-2.85	0.64	0.62	0.65	0.64	0.62
H3021C01	-3.02	-1.34	-2.13	-2.05	-2.25	0.66	0.64	0.68	0.66	0.67
H3039C07	-1.35	-0.70	-0.73	-0.87	-1.03	0.28	0.27	0.28	0.28	0.27
H3016D01	-2.04	-0.31	-1.03	-1.38	-1.25	0.52	0.50	0.52	0.52	0.50
H3047C02	-3.62	-1.66	-2.61	-2.77	-2.85	0.71	0.69	0.72	0.71	0.69
glutamic pyruvate transaminase (alanine aminotransferase) 2	-3.82	-2.06	-2.70	-2.98	-3.01	0.68	0.66	0.69	0.68	0.66
H3032D10	-3.29	-1.91	-2.62	-2.94	-2.86	0.56	0.55	0.57	0.56	0.54
H3043D10	-1.77	-0.77	-1.16	-1.00	-0.71	0.31	0.30	0.30	0.30	0.29
triple functional domain (PTPRF interacting)	-1.01	-0.48	-0.86	-0.84	-0.86	0.21	0.21	0.21	0.21	0.20
RAN guanine nucleotide release factor	-1.39	-0.81	-0.66	-0.91	-0.84	0.24	0.23	0.24	0.24	0.23
ARP1 actin-related protein 1 homolog A (yeast)	-3.23	-1.59	-2.39	-2.59	-2.85	0.53	0.52	0.54	0.53	0.52
small glutamine-rich tetratricopeptide repeat	-4.38	-2.21	-3.10	-3.17	-3.26	0.80	0.77	0.81	0.80	0.77

(TPR)-containing, alpha										
H3037C11	-4.39	-1.20	-2.19	-2.07	-2.66	0.88	0.86	0.89	0.88	0.85
dehydrogenase/reductase (SDR family) X chromosome	-2.30	-1.35	-1.10	-1.61	-1.44	0.32	0.31	0.32	0.32	0.31
Hspb associated protein 1	-3.75	-1.87	-2.56	-2.76	-3.00	0.68	0.66	0.69	0.68	0.66
Rho-associated coiled-coil forming kinase 1	-2.45	-1.44	-1.46	-2.03	-1.91	0.39	0.38	0.40	0.41	0.38
H3033D12	-3.74	-1.42	-2.68	-2.85	-2.83	0.67	0.65	0.68	0.67	0.65
H3123D06	-2.43	-0.75	-1.85	-1.85	-2.00	0.57	0.55	0.58	0.57	0.55
H3048E01	-3.28	-1.10	-1.96	-1.65	-2.19	0.67	0.65	0.68	0.67	0.65
heterogeneous nuclear ribonucleoprotein D-like	-2.18	-0.85	-1.44	-1.37	-1.81	0.42	0.41	0.43	0.42	0.41
DNA segment, Chr 14, ERATO Doi 436, expressed	-3.38	-1.08	-1.77	-1.82	-2.40	0.68	0.66	0.69	0.68	0.66
UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 1	-2.05	-1.21	-0.98	-1.53	-1.52	0.46	0.44	0.46	0.45	0.44
H3116F08	-3.73	-1.74	-2.47	-2.74	-3.05	0.68	0.66	0.69	0.68	0.65
metallothionein 2	-2.27	-1.19	-1.00	-1.47	-1.36	0.45	0.44	0.46	0.45	0.44
H3020E05	-2.30	-1.54	-1.02	-2.19	-1.67	0.64	0.63	0.65	0.64	0.62
histidyl-tRNA synthetase-like	-4.63	-3.08	-1.83	-3.72	-2.87	0.73	0.71	0.74	0.73	0.70
H3060E06	-1.71	-1.15	-1.45	-1.71	-1.46	0.28	0.27	0.28	0.28	0.27
H3072E06	-2.03	-1.35	-1.31	-1.89	-1.58	0.37	0.36	0.38	0.37	0.36
gene model 179, (NCBI)	-2.88	-1.93	-2.67	-2.66	-2.51	0.67	0.65	0.68	0.67	0.65
myosin 1H	-3.72	-2.10	-2.75	-3.04	-3.07	0.74	0.72	0.76	0.74	0.72
H3152H01	-1.77	-0.80	-0.96	-1.32	-1.32	0.26	0.25	0.27	0.26	0.25
ankyrin repeat domain 17	-3.05	-1.87	-2.47	-2.76	-2.54	0.61	0.60	0.62	0.61	0.59
H3036H08	-3.42	-1.82	-2.52	-2.92	-3.03	0.62	0.60	0.63	0.62	0.60
H3035G10	-3.98	-1.96	-2.85	-3.00	-3.03	0.73	0.71	0.74	0.73	0.71
H3116G12	-3.66	-1.71	-2.73	-2.67	-3.03	0.60	0.58	0.61	0.60	0.58

Cluster 4 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
H3054B02	-1.76	-1.22	-0.40	-1.50	-1.56	0.41	0.40	0.41	0.41	0.39
RIKEN cDNA 1110060F11 gene	-2.42	-1.57	-0.83	-1.26	-0.74	0.54	0.53	0.55	0.54	0.53
arginine-rich, mutated in early stage tumors	-1.78	-1.27	-0.11	-0.76	-0.96	0.42	0.41	0.42	0.42	0.40
ribosomal protein L4	-1.77	-1.23	-0.40	-1.20	-0.91	0.25	0.24	0.25	0.25	0.24
glomulin, FKBP associated protein	-1.38	-0.72	-0.49	-0.82	-0.56	0.22	0.21	0.22	0.22	0.21
RIKEN cDNA 1110036D12 gene	-1.50	-1.21	-0.66	-1.30	-0.46	0.37	0.36	0.37	0.37	0.36
ubiquitin specific protease 7	-1.82	-0.88	-0.73	-1.62	-0.30	0.41	0.40	0.41	0.41	0.39
zinc finger, CCHC domain containing 10	-2.04	-0.91	-0.58	-0.46	-0.52	0.33	0.33	0.33	0.33	0.32
phospholipase A2, group IVC (cytosolic, calcium-independent)	-2.01	-1.55	-0.65	-1.26	-1.12	0.32	0.32	0.33	0.32	0.31
cellular repressor of E1A-stimulated genes 1	-2.97	-1.65	-0.82	-1.55	-0.81	0.78	0.76	0.80	0.78	0.76
cDNA sequence BC038479	-2.99	-1.74	-1.48	-1.97	-1.62	0.41	0.40	0.41	0.41	0.40
mannosidase, beta A, lysosomal	-1.21	-0.74	-0.53	-0.82	-0.61	0.23	0.22	0.23	0.22	0.22
H3075C11	-2.42	-1.87	-0.74	-1.59	-1.36	0.35	0.34	0.36	0.35	0.34
H3011D05	-2.49	-1.48	-0.47	-1.20	-1.15	0.29	0.28	0.29	0.29	0.28
hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), beta subunit	-2.54	-1.81	-0.47	-1.49	-0.94	0.57	0.55	0.58	0.57	0.55
DNA segment, Chr 14, ERATO Doi 500, expressed	-1.15	-0.64	-0.39	-0.67	-0.49	0.18	0.17	0.18	0.18	0.18
armadillo repeat containing, X-linked 3	-1.38	-0.81	-0.08	-1.03	-0.80	0.37	0.36	0.37	0.37	0.36
RIKEN cDNA 1200009C21 gene	-0.83	-0.72	-0.16	-0.61	-0.42	0.19	0.18	0.19	0.19	0.18
RIKEN full-length enriched library, clone:6720415B15	-1.44	-1.00	-0.57	-1.04	-0.65	0.32	0.32	0.34	0.32	0.31
dehydrogenase/reductase (SDR family) member 1	-2.53	-0.75	-0.44	-1.35	-0.93	0.54	0.53	0.55	0.54	0.55
H3041F01	-1.58	-1.32	-0.36	-1.16	-0.96	0.35	0.34	0.36	0.35	0.34
ribosome binding protein 1	-0.88	-0.74	-0.12	-0.55	-0.71	0.21	0.21	0.22	0.21	0.21
H3091F09	-1.37	-0.92	-0.14	-0.54	-0.41	0.22	0.21	0.22	0.21	0.21
ribosomal protein S20	-1.43	-0.79	-0.10	-0.77	-0.67	0.29	0.28	0.29	0.29	0.28
H3121F11	-1.36	-1.03	-0.15	-0.65	-0.76	0.29	0.28	0.29	0.29	0.28
expressed sequence AI848100	-2.59	-1.42	-0.45	-0.88	-1.22	0.33	0.32	0.33	0.33	0.31

splicing factor, arginine/serine-rich 7	-1.64	-0.83	-0.31	-0.37	-0.41	0.35	0.35	0.35	0.36	0.34
RIKEN cDNA 0610009E20 gene	-1.59	-1.58	-0.07	-1.06	-0.86	0.26	0.26	0.28	0.27	0.25
H3082H10	-2.10	-0.90	-0.36	-0.85	-0.69	0.49	0.47	0.50	0.49	0.47
RIKEN cDNA 1110001A05 gene	-2.22	-0.80	-0.03	-0.71	-0.54	0.55	0.54	0.56	0.55	0.53
H3063G06	-1.19	-0.80	-0.17	-0.79	-0.53	0.26	0.26	0.27	0.26	0.25
H3064G06	-1.44	-0.30	-0.21	-0.40	-0.77	0.30	0.29	0.30	0.30	0.29
H3012H06	-1.09	-0.76	-0.25	-0.68	-0.61	0.23	0.22	0.23	0.23	0.22
H3020H12	-1.73	-1.25	-0.08	-0.82	-0.76	0.28	0.27	0.28	0.28	0.27
H3023H06	-1.61	-1.15	-0.22	-0.76	-0.68	0.32	0.33	0.33	0.33	0.31
H3065H06	-1.78	-0.70	-0.20	-1.18	-0.61	0.39	0.37	0.39	0.39	0.37

Cluster 5 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
Mus musculus LOC381030 (LOC381030), mRNA	-1.81	-1.58	-0.95	-1.54	-1.02	0.44	0.42	0.44	0.44	0.43
gene model 1024, (NCBI)	-2.21	-1.77	-1.68	-2.13	-1.80	0.28	0.27	0.28	0.28	0.27
H3150B03	-0.48	-1.52	-0.71	-1.49	-0.57	0.36	0.38	0.36	0.36	0.34
H3012B10	-0.14	-1.10	0.10	-0.98	-0.50	0.32	0.30	0.31	0.31	0.30
RIKEN cDNA B230106I24 gene	-1.68	-2.40	-1.05	-2.18	-1.52	0.41	0.40	0.42	0.41	0.40
H3134A05	-1.76	-1.55	-0.85	-1.38	-1.10	0.31	0.30	0.31	0.31	0.30
PHD finger protein 16	-0.91	-0.93	-0.34	-1.02	-0.62	0.23	0.22	0.23	0.23	0.22
RIKEN cDNA 9230112O05 gene	-1.57	-1.74	-1.22	-1.79	-1.50	0.26	0.25	0.26	0.26	0.25
RIKEN cDNA 5830411E10 gene	-2.06	-1.57	-1.75	-2.24	-1.22	0.61	0.58	0.61	0.60	0.58
H3043C08	-1.87	-1.72	-1.41	-1.82	-1.32	0.35	0.34	0.36	0.35	0.34
eukaryotic translation elongation factor 1 alpha 1	-2.65	-1.99	-1.67	-2.22	-1.99	0.47	0.46	0.48	0.47	0.46
H3068C09	-5.39	-5.38	-5.09	-5.15	-4.53	0.35	0.35	0.36	0.35	0.34
nucleolar protein 8	-1.30	-1.04	-0.94	-0.87	-0.81	0.26	0.25	0.26	0.26	0.25
pyruvate kinase, muscle	-1.15	-1.13	-0.75	-1.41	-0.73	0.21	0.20	0.21	0.21	0.21
general transcription factor III A	-1.50	-1.34	-0.89	-1.60	-1.00	0.31	0.30	0.31	0.31	0.30
H3041C05	-1.54	-1.48	-0.92	-1.41	-1.05	0.28	0.28	0.29	0.28	0.27
microtubule-associated protein 7	-0.70	-1.23	-0.58	-1.09	-0.67	0.21	0.21	0.22	0.21	0.21
H3135F01	-1.73	-1.72	-1.68	-1.88	-1.50	0.31	0.30	0.31	0.31	0.30
general transcription factor II H, polypeptide 4	-1.90	-2.12	-1.07	-1.52	-1.39	0.34	0.33	0.34	0.35	0.32
splicing factor 3b, subunit 2	-2.64	-2.20	-2.56	-2.33	-2.03	0.29	0.28	0.30	0.29	0.28
ATPase, H ⁺ transporting, V1 subunit E isoform 1	-1.34	-1.15	-0.40	-1.27	-1.00	0.28	0.27	0.28	0.28	0.27
caspase 8 associated protein 2	-1.67	-1.91	-1.05	-1.79	-1.34	0.36	0.35	0.37	0.36	0.35
C-type (calcium dependent, carbohydrate recognition domain) lectin, superfamily member 8	-1.53	-1.09	-1.22	-1.21	-1.03	0.28	0.28	0.29	0.28	0.27
gene rich cluster, C2f gene	-2.20	-2.06	-2.04	-2.01	-1.58	0.33	0.32	0.33	0.33	0.32
H3076F06	-2.60	-2.30	-1.74	-2.61	-2.15	0.47	0.46	0.48	0.47	0.46
hemoglobin alpha, adult chain 1	-2.73	-2.47	-1.98	-2.60	-2.30	0.46	0.44	0.46	0.46	0.44
DNA segment, Chr 7, ERATO Doi 743, expressed	-2.23	-2.02	-1.09	-2.04	-1.40	0.47	0.46	0.48	0.47	0.45
nucleoporin 153	-1.06	-0.92	-0.58	-1.25	-0.86	0.28	0.26	0.28	0.27	0.27
homeo box B1	-1.45	-1.35	-1.13	-1.16	-0.90	0.23	0.23	0.24	0.24	0.22
SLIT-ROBO Rho GTPase	-1.43	-1.21	-1.05	-1.37	-1.00	0.28	0.27	0.29	0.28	0.27

activating protein 2										
bone marrow stromal cell antigen 2	-0.94	-1.01	-0.39	-1.19	-1.05	0.28	0.27	0.28	0.28	0.27
Z3002G10	-1.23	-2.00	-1.00	-1.97	-1.16	0.31	0.30	0.31	0.31	0.31
H3003G05	-1.19	-1.09	-0.52	-1.09	-0.77	0.22	0.22	0.23	0.23	0.22
H3061G12	-2.48	-2.09	-2.20	-2.08	-1.84	0.33	0.32	0.33	0.33	0.32
H3105G12	-1.74	-1.73	-0.88	-1.36	-1.17	0.24	0.23	0.24	0.23	0.23
transcriptional regulator protein	-1.29	-0.93	-0.72	-1.27	-0.76	0.29	0.28	0.30	0.30	0.28

Cluster 6 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
RIKEN cDNA 2410195B05 gene	2.01	3.31	1.96	1.37	1.72	0.65	0.64	0.66	0.65	0.64
ubiquitin specific protease 9, X chromosome	0.76	0.75	-0.38	-0.37	-0.08	0.29	0.28	0.29	0.29	0.28
polymerase (RNA) II (DNA directed) polypeptide G	0.48	1.12	0.34	0.55	0.06	0.24	0.23	0.24	0.24	0.23
H3144B02	1.08	1.63	0.16	0.01	0.22	0.32	0.30	0.31	0.31	0.30
H3085A03	1.43	2.03	0.73	-0.03	0.84	0.24	0.24	0.25	0.24	0.23
H3150A06	-0.29	1.38	0.31	0.01	0.29	0.35	0.32	0.33	0.32	0.31
H3135C07	1.47	1.93	0.32	-0.23	0.21	0.46	0.45	0.47	0.46	0.44
Mus musculus 2 days neonate thymus thymic cells cDNA, RIKEN full-length enriched library, clone:E430005D23 product:hypothetical protein, full insert sequence.	1.62	1.58	0.48	0.22	0.58	0.41	0.40	0.42	0.41	0.40
minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe) (S. cerevisiae)	0.67	2.03	0.27	-0.11	0.21	0.43	0.41	0.43	0.43	0.41
H3154D08	1.15	1.30	0.29	0.12	0.49	0.30	0.27	0.29	0.27	0.27
H3135C09	1.70	2.40	0.62	0.15	0.58	0.34	0.33	0.34	0.34	0.33
H3061C10	0.11	0.83	0.02	0.23	0.16	0.18	0.16	0.17	0.17	0.16
RNA binding motif protein 16	0.94	1.88	0.05	-0.19	0.31	0.41	0.40	0.42	0.41	0.40
upstream regulatory element binding protein 1e	0.97	1.33	0.26	-0.05	0.16	0.34	0.33	0.34	0.34	0.32
Kruppel-like factor 7 (ubiquitous)	0.97	1.06	0.27	-0.06	0.04	0.23	0.23	0.23	0.24	0.22
RIKEN cDNA 1700013H19 gene	0.97	1.30	0.37	0.29	0.10	0.29	0.29	0.29	0.29	0.29
LIM domain only 7	0.85	1.60	0.40	0.48	0.43	0.32	0.31	0.32	0.32	0.31
H3037F07	0.49	1.32	-0.13	0.01	0.13	0.32	0.32	0.33	0.32	0.31
Z3002F07	0.58	1.88	0.28	0.32	0.08	0.30	0.30	0.30	0.30	0.30
synovial sarcoma translocation gene on chromosome 18-like 1	-0.02	1.26	-0.05	-0.59	-0.17	0.31	0.30	0.31	0.31	0.30
H3151F02	0.37	1.35	-0.06	-0.87	-0.41	0.52	0.47	0.49	0.47	0.46
H3009E10	0.89	1.58	0.39	0.15	0.48	0.26	0.26	0.27	0.26	0.25
RIKEN cDNA 2210010C17 gene	1.42	2.08	0.47	0.51	0.66	0.40	0.39	0.41	0.40	0.39
H3132F10	1.67	1.85	1.03	0.60	0.87	0.38	0.37	0.39	0.38	0.37
protein tyrosine phosphatase, receptor type,	4.35	4.85	3.01	1.38	2.33	0.56	0.54	0.57	0.56	0.54

E										
H3137F06	1.00	1.69	0.31	-0.60	0.26	0.51	0.50	0.52	0.51	0.50
H3055H07	1.90	1.91	0.50	-0.05	0.82	0.59	0.57	0.60	0.59	0.57
H3132H02	0.54	1.45	0.69	0.56	0.73	0.29	0.28	0.29	0.29	0.28
H3045G11	0.67	1.69	0.62	0.87	0.68	0.24	0.23	0.24	0.24	0.23
ferritin light chain 1	0.21	2.02	0.79	0.08	-0.12	0.20	0.20	0.21	0.20	0.20
ankyrin repeat and SOCS box-containing protein 1	1.27	1.70	0.45	0.46	0.84	0.33	0.33	0.34	0.33	0.32
Similar to phosphatidylserine decarboxylase.	0.77	0.93	0.36	-0.13	-0.35	0.24	0.23	0.24	0.24	0.23

Cluster 7 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
H3071B02	-1.72	-1.16	-0.46	-0.02	-0.15	0.39	0.38	0.39	0.39	0.38
nucleoporin 37	-1.07	-1.09	-0.08	0.29	0.36	0.37	0.36	0.37	0.37	0.36
H3005D07	-1.24	-0.51	0.54	0.24	0.44	0.33	0.32	0.33	0.33	0.31
abhydrolase domain containing 3	-1.18	-0.92	-0.08	-0.27	-0.06	0.30	0.29	0.30	0.30	0.29
platelet derived growth factor receptor, beta polypeptide	-1.15	-0.99	-0.04	-0.42	-0.18	0.20	0.20	0.20	0.20	0.19
capping protein (actin filament) muscle Z-line, alpha 1	-1.28	-0.60	0.37	0.14	0.24	0.29	0.28	0.29	0.29	0.28
H3057C03	-2.64	-1.33	-0.02	-1.00	-0.42	0.60	0.58	0.61	0.60	0.58
H3088D03	-2.35	-1.67	-0.02	-0.67	-0.23	0.65	0.64	0.66	0.65	0.63
FK506 binding protein 8	-1.58	-0.97	-0.18	-0.40	-0.21	0.35	0.34	0.35	0.35	0.34
ninein	-1.08	-0.70	0.36	-0.68	0.04	0.32	0.30	0.31	0.30	0.30
proteasome (prosome, macropain) subunit, alpha type 6	-1.13	-0.83	-0.19	-0.23	-0.07	0.22	0.22	0.23	0.22	0.22
H3048D12	-1.77	-1.00	-0.13	-0.67	-0.04	0.42	0.41	0.43	0.42	0.41
RIKEN full-length enriched library, clone:D930002I12	-1.73	-1.08	-0.27	-0.63	-0.06	0.25	0.24	0.25	0.25	0.24
molybdenum cofactor synthesis 3	-2.09	-0.82	0.73	0.52	0.72	0.33	0.32	0.34	0.33	0.32
ubiquinol-cytochrome c reductase hinge protein	-0.80	-0.76	-0.09	-0.28	0.15	0.22	0.21	0.22	0.21	0.22
hemoglobin alpha, adult chain 1	-1.68	-0.93	0.00	-0.27	0.02	0.36	0.35	0.37	0.36	0.35
ribosomal protein L3	-0.99	-1.04	0.29	-0.08	0.15	0.33	0.32	0.33	0.33	0.32
brain-specific angiogenesis inhibitor 1-associated protein 2	-0.77	-0.38	0.33	-0.11	0.20	0.18	0.18	0.19	0.18	0.17
H3098F03	-1.18	-0.50	0.32	0.33	0.08	0.24	0.23	0.24	0.24	0.23
H3002E04	-1.00	-0.48	0.58	0.14	0.43	0.32	0.31	0.33	0.32	0.31
RIKEN cDNA 2410141M05 gene	-1.88	-0.65	0.17	0.19	-0.02	0.41	0.40	0.42	0.41	0.40
H3046E04	-1.44	-0.66	0.10	0.24	0.23	0.30	0.29	0.30	0.30	0.29
basonuclin 2	-1.65	-1.11	-0.54	-0.48	0.09	0.40	0.38	0.40	0.40	0.38
H3096F10	-0.87	-0.26	0.58	0.48	0.37	0.31	0.30	0.32	0.31	0.30
3-phosphoglycerate dehydrogenase	-2.06	-0.63	0.54	0.90	0.58	0.50	0.48	0.50	0.49	0.47
H3103F11	-0.96	-0.16	0.40	0.65	0.51	0.30	0.29	0.30	0.30	0.29
dual specificity phosphatase 19	-1.86	-0.79	0.25	0.58	0.85	0.42	0.41	0.43	0.42	0.41
WW domain containing E3 ubiquitin protein ligase 2	-1.20	-0.82	-0.12	0.10	-0.17	0.24	0.23	0.24	0.24	0.23

H3104H07	-1.54	-0.93	0.15	0.10	0.01	0.40	0.39	0.41	0.40	0.39
H3050G03	-1.45	-0.96	0.29	0.19	0.07	0.34	0.33	0.35	0.34	0.33
apolipoprotein A-IV	-0.72	-0.66	0.00	0.06	0.12	0.20	0.18	0.19	0.18	0.18
RIKEN cDNA 4930579C15 gene	-1.18	-0.45	-0.08	-0.18	-0.01	0.26	0.26	0.27	0.26	0.25
H3102G03	-0.73	-0.93	0.03	-0.10	0.06	0.14	0.14	0.14	0.14	0.13
myosin IC	-3.03	-1.90	0.68	-0.96	0.04	0.67	0.65	0.68	0.67	0.65

Cluster 8 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
H3017B01	-1.80	-1.90	-0.62	-0.21	-0.24	0.57	0.55	0.58	0.57	0.55
3-phosphoglycerate dehydrogenase	-1.43	-1.00	-0.80	-0.69	-1.05	0.29	0.28	0.30	0.29	0.28
H3063B11	-1.70	-2.35	-1.89	-1.14	-1.00	0.48	0.46	0.48	0.48	0.46
protein kinase C, zeta	-2.35	-1.62	-1.02	-1.22	-0.91	0.44	0.43	0.45	0.44	0.42
H3039C08	-1.69	-1.58	-0.79	-0.80	-0.80	0.30	0.29	0.30	0.30	0.29
tankyrase 1 binding protein 1	-1.12	-1.42	-0.32	-0.80	-0.51	0.17	0.17	0.17	0.17	0.17
RIKEN cDNA 9430034D17 gene	-1.80	-1.39	-0.58	-1.04	-0.69	0.31	0.31	0.32	0.31	0.30
H3094C04	-1.12	-1.11	-0.59	-0.89	-0.40	0.23	0.23	0.24	0.23	0.23
RIKEN cDNA 4921506I22 gene	-2.01	-1.41	-1.24	-0.77	-1.00	0.43	0.41	0.43	0.43	0.41
isocitrate dehydrogenase 3 (NAD+) alpha	-0.86	-1.05	-0.15	-0.56	-0.49	0.20	0.20	0.21	0.20	0.20
macrophage erythroblast attacher	-2.36	-1.77	-1.18	-1.40	-1.27	0.49	0.48	0.50	0.49	0.48
RIKEN cDNA 1200014M14 gene	-2.15	-0.89	-1.07	-0.21	-0.42	0.42	0.40	0.41	0.40	0.39
cullin 2	-2.99	-2.09	-1.51	-1.40	-1.39	0.25	0.24	0.26	0.25	0.24
H3035D05	-2.71	-2.52	-0.86	-0.84	-1.41	0.58	0.58	0.59	0.58	0.55
H3042E02	-2.26	-1.50	-1.02	-0.77	-1.12	0.36	0.35	0.37	0.36	0.35
H3037E03	-1.46	-0.88	-0.61	-0.32	-0.54	0.26	0.26	0.27	0.26	0.25
thyroid hormone receptor interactor 12	-0.77	-1.10	-0.79	-0.62	-0.36	0.21	0.21	0.21	0.21	0.20
RIKEN cDNA 3930401E15 gene	-1.45	-1.46	-0.90	-1.04	-0.54	0.20	0.19	0.20	0.20	0.19
ribosomal protein L6	-1.28	-1.13	-0.48	-0.75	-0.49	0.29	0.29	0.30	0.29	0.29
ubiquitin-conjugating enzyme E2L 3	-1.32	-1.12	-0.33	-0.60	-0.57	0.28	0.27	0.28	0.28	0.27
H3016F10	-1.86	-1.46	-0.86	-1.12	-0.90	0.37	0.36	0.37	0.37	0.36
H3054F04	-2.34	-1.83	-1.24	-0.57	-0.40	0.46	0.44	0.46	0.46	0.44
H3039H07	-2.25	-1.96	-1.20	-0.79	-0.93	0.35	0.33	0.35	0.34	0.33
nucleosome assembly protein 1-like 1	-0.78	-1.00	-0.24	-0.87	-0.22	0.25	0.24	0.25	0.25	0.26
H3018G09	-1.37	-1.18	-0.49	-0.68	-0.29	0.25	0.25	0.26	0.26	0.24
expressed sequence AW544981	-2.03	-1.08	-0.80	-0.34	-0.36	0.36	0.35	0.36	0.36	0.36
H3018G10	-1.09	-1.25	-0.65	-0.67	-0.52	0.23	0.23	0.23	0.23	0.22
glutamate-ammonia ligase (glutamine synthase)	-1.91	-1.81	-1.04	-0.35	-0.71	0.32	0.31	0.32	0.32	0.31
RIKEN cDNA 1200013F24 gene	-0.93	-0.99	-0.06	-0.37	-0.32	0.23	0.22	0.23	0.23	0.22
H3030G06	-2.90	-1.88	-1.51	-1.47	-1.45	0.40	0.39	0.41	0.40	0.39

Cluster 9 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
H3137A08	-1.72	-1.33	-1.42	-2.12	-1.51	0.51	0.50	0.52	0.51	0.50
RIKEN cDNA A430089I19 gene	-2.83	-2.65	-2.92	-4.13	-3.43	0.47	0.45	0.47	0.47	0.45
RIKEN cDNA 1110019J04 gene	-2.12	-2.72	-2.13	-3.23	-3.05	0.47	0.45	0.47	0.47	0.45
elongation factor Tu GTP binding domain containing 1	-1.43	-1.45	-1.75	-1.72	-1.34	0.24	0.24	0.25	0.25	0.23
RIKEN cDNA 6030443O07 gene	-3.34	-4.25	-4.86	-4.39	-3.74	0.38	0.37	0.39	0.38	0.37
pyruvate carboxylase	-1.73	-1.38	-1.30	-1.97	-1.35	0.41	0.41	0.42	0.42	0.40
H3151A09	-0.46	-1.30	-1.19	-0.92	-0.61	0.26	0.26	0.27	0.26	0.26
H3127A12	-1.73	-1.93	-2.07	-2.66	-2.81	0.32	0.31	0.32	0.32	0.31
H3032C08	-0.90	-1.77	-1.43	-1.54	-1.20	0.38	0.35	0.37	0.36	0.35
RIKEN full-length enriched library, clone:E230012I16	-1.24	-0.95	-0.98	-1.52	-1.23	0.35	0.34	0.35	0.35	0.34
H3128D10	-1.22	-2.02	-3.37	-2.75	-2.20	0.58	0.56	0.59	0.58	0.56
H3147C05	-2.23	-1.62	-1.71	-2.43	-2.04	0.57	0.56	0.58	0.57	0.55
H3105C12	-1.48	-2.16	-2.18	-2.28	-1.79	0.25	0.24	0.25	0.25	0.24
H3147C06	-1.36	-1.24	-1.35	-2.34	-1.59	0.39	0.38	0.39	0.38	0.37
H3139D06	-1.90	-1.41	-1.21	-2.08	-1.60	0.36	0.36	0.35	0.35	0.34
zinc finger, DHHC domain containing 7	-1.29	-0.77	-0.95	-1.42	-1.37	0.29	0.28	0.29	0.29	0.28
H3012E09	-2.44	-1.68	-1.53	-3.35	-2.17	0.71	0.70	0.73	0.71	0.69
RIKEN cDNA 2310021P13 gene	-2.26	-3.22	-3.42	-3.35	-2.95	0.37	0.33	0.35	0.34	0.33
H3030E11	-1.16	-1.19	-1.07	-1.55	-1.39	0.32	0.31	0.32	0.32	0.31
H3037E11	-1.21	-1.11	-1.58	-1.31	-1.41	0.36	0.35	0.37	0.36	0.35
H3037F11	-1.72	-1.24	-1.86	-1.57	-1.81	0.41	0.40	0.42	0.41	0.40
RIKEN cDNA 2310003H01 gene	-2.09	-2.38	-2.52	-3.30	-2.81	0.37	0.36	0.38	0.37	0.36
RIKEN cDNA 2810406C15 gene	-0.77	-1.12	-0.61	-1.23	-1.36	0.27	0.26	0.27	0.27	0.26
H3114F12	-1.51	-2.02	-1.96	-2.60	-2.16	0.40	0.39	0.41	0.40	0.39
H3061H01	-1.96	-1.42	-1.70	-3.06	-1.92	0.41	0.41	0.42	0.42	0.41
solute carrier organic anion transporter family, member 2a1	-1.43	-1.17	-2.52	-2.47	-2.05	0.30	0.29	0.30	0.30	0.30
H3141H01	-1.62	-1.28	-1.79	-1.68	-1.91	0.33	0.32	0.33	0.33	0.32
RIKEN cDNA 3100004P22 gene	-1.36	-1.92	-1.76	-2.86	-2.05	0.47	0.46	0.48	0.47	0.45
MYST histone acetyltransferase (monocytic leukemia) 3	-0.85	-0.97	-1.14	-1.77	-1.36	0.34	0.33	0.35	0.34	0.33
H3121H02	-1.66	-0.89	-2.04	-2.09	-1.70	0.35	0.34	0.35	0.35	0.34

RIKEN cDNA 4930473A02 gene	-0.89	-0.73	-0.66	-0.77	-1.52	0.30	0.29	0.30	0.30	0.29
H3011H05	-0.75	-1.11	-0.65	-1.12	-1.19	0.26	0.25	0.26	0.26	0.25
neighbor of Brca1 gene 1	-1.56	-1.18	-1.23	-2.04	-1.98	0.37	0.36	0.38	0.37	0.36
Glutaminase	-1.35	-0.94	-1.80	-1.44	-1.34	0.26	0.26	0.27	0.27	0.26

Cluster 10 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
kinesin family member 2C	-0.59	-0.60	-0.20	-1.42	-0.34	0.29	0.28	0.30	0.30	0.28
H3011B10	-0.03	-0.73	-0.45	-1.32	-1.21	0.33	0.32	0.34	0.33	0.32
H3021B10	-0.75	-0.35	-0.06	-1.22	-0.67	0.28	0.27	0.28	0.28	0.27
SMC2 structural maintenance of chromosomes 2-like 1 (yeast)	-0.85	0.13	-0.50	-1.79	-1.13	0.40	0.38	0.40	0.40	0.38
GATA binding protein 6	-0.65	-0.76	-0.66	-1.93	-1.45	0.33	0.32	0.34	0.33	0.32
signal recognition particle 9	-0.62	0.02	-0.27	-1.45	-0.62	0.34	0.33	0.35	0.34	0.33
H3038A12	-0.45	-0.87	-0.82	-1.39	-1.28	0.28	0.28	0.29	0.28	0.28
H3046B12	-0.64	-0.34	-0.53	-2.15	-1.64	0.51	0.50	0.52	0.51	0.49
H3143B12	-0.31	0.36	-1.29	-1.57	-1.41	0.43	0.42	0.43	0.43	0.41
H3146B06	0.24	-0.19	-0.42	-1.99	-1.48	0.54	0.53	0.55	0.54	0.53
H3091C01	-1.18	-0.84	-0.29	-2.51	-0.93	0.65	0.63	0.66	0.67	0.65
oxidative-stress responsive 1	-1.04	-0.12	-0.66	-1.16	-0.51	0.29	0.28	0.30	0.29	0.28
RAB18, member RAS oncogene family	-0.89	-0.36	-0.02	-1.39	-0.91	0.32	0.32	0.33	0.32	0.31
expressed sequence AA517853	-0.50	0.67	-0.52	-1.55	-0.80	0.51	0.49	0.51	0.51	0.49
interleukin 3 receptor, alpha chain	-0.70	0.29	-1.06	-1.18	-1.24	0.30	0.29	0.30	0.30	0.29
thymoma viral proto-oncogene 1	-0.80	-0.65	-0.42	-1.57	-0.69	0.31	0.30	0.32	0.31	0.30
RIKEN cDNA 2610510J17 gene	-0.89	-0.97	-1.05	-2.45	-1.73	0.35	0.34	0.36	0.35	0.34
RIKEN cDNA 1110054O05 gene	-0.12	-0.63	-1.10	-1.68	-1.05	0.34	0.33	0.35	0.34	0.33
G protein-coupled receptor 116	-2.29	-0.10	-0.53	-2.40	-1.60	0.88	0.87	0.91	0.88	0.85
anaphase promoting complex subunit 4	-1.00	-0.80	-0.97	-2.37	-1.58	0.48	0.47	0.49	0.48	0.46

Cluster 11 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
H3073A08	-0.97	0.36	1.48	0.97	0.78	0.55	0.53	0.56	0.55	0.53
RAB23, member RAS oncogene family	-0.51	1.09	2.89	2.13	2.15	0.59	0.57	0.60	0.59	0.57
heterogeneous nuclear ribonucleoprotein M	-0.99	0.27	0.43	0.30	0.80	0.37	0.36	0.37	0.37	0.35
RIKEN cDNA 1110019N10 gene	-0.19	0.07	1.06	0.95	0.67	0.27	0.26	0.27	0.27	0.26
H3048D05	-0.68	0.02	0.71	0.62	1.04	0.25	0.25	0.25	0.25	0.24
H3115E08	-0.62	-0.26	0.86	0.59	0.67	0.33	0.31	0.32	0.32	0.31
hexosaminidase B	-0.63	0.29	0.71	0.66	1.03	0.35	0.34	0.35	0.35	0.33
RNA binding motif protein 3	-1.70	-0.17	0.89	1.10	0.21	0.39	0.38	0.40	0.39	0.38
H3118F10	-0.42	0.22	1.11	0.81	1.24	0.36	0.35	0.37	0.36	0.35
myosin IE	-0.83	0.12	1.53	1.11	0.80	0.54	0.52	0.55	0.54	0.52
H3010F12	0.22	0.12	0.58	0.46	0.51	0.10	0.10	0.10	0.10	0.09
DNA segment, Chr 8, ERATO Doi 82, expressed	-0.28	-0.34	0.42	0.65	0.72	0.21	0.20	0.20	0.20	0.19
H3034G01	0.05	0.33	1.35	0.60	1.01	0.32	0.33	0.33	0.32	0.31
CREB binding protein	-0.76	-0.63	1.55	1.40	0.82	0.50	0.48	0.50	0.50	0.48
placental specific protein 1	-0.07	0.31	0.52	0.52	0.62	0.14	0.13	0.14	0.14	0.13
H3034G09	-1.15	0.32	0.93	1.04	1.31	0.49	0.48	0.50	0.49	0.48
tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta polypeptide	-0.94	-0.56	1.00	0.34	0.17	0.42	0.40	0.42	0.42	0.40
Dip3 beta	-0.94	0.08	1.18	0.57	0.71	0.38	0.37	0.38	0.38	0.37
RIKEN cDNA 2610019P18 gene	-0.81	0.56	0.81	0.90	1.02	0.38	0.37	0.38	0.38	0.36
H3024H11	-1.26	0.68	1.70	1.52	1.39	0.39	0.36	0.39	0.37	0.36

Cluster 12 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
coiled-coil domain containing 2	-0.44	0.49	1.73	0.42	0.65	0.47	0.46	0.48	0.48	0.48
H3030A03	-2.54	0.77	1.46	-0.02	0.47	0.95	0.93	0.97	0.95	0.92
pleckstrin homology domain containing, family H (with MyTH4 domain) member 1	-0.82	1.59	2.74	1.22	1.25	0.75	0.72	0.76	0.75	0.72
H3011A04	-1.97	-0.36	0.58	-0.50	-0.44	0.44	0.43	0.44	0.43	0.42
acid phosphatase 1, soluble	-2.25	-0.34	0.79	-0.02	0.15	0.45	0.44	0.46	0.45	0.44
H3030A06	-0.36	1.35	2.12	0.75	1.06	0.38	0.37	0.39	0.38	0.36
H3064B12	-0.49	-0.01	2.50	0.42	0.15	0.68	0.66	0.69	0.68	0.66
H3010D07	-2.01	-0.80	0.91	-0.59	-0.38	0.51	0.49	0.52	0.51	0.49
inhibitor of growth family, member 1	-2.06	-0.26	0.43	0.04	-0.01	0.55	0.54	0.56	0.55	0.53
ribosomal protein S15a	-1.70	-0.07	1.40	0.09	0.14	0.46	0.44	0.46	0.46	0.44
H3155C11	-1.09	0.38	0.66	0.30	0.02	0.37	0.36	0.38	0.37	0.36
Wiskott-Aldrich syndrome protein interacting protein	-1.07	-0.23	0.07	-0.14	0.29	0.28	0.28	0.28	0.28	0.27
H3109D12	-0.48	0.28	1.02	0.03	0.16	0.28	0.27	0.28	0.28	0.27
H3056E02	-1.47	0.09	0.96	-0.36	-0.80	0.54	0.52	0.55	0.54	0.54
platelet-activating factor acetylhydrolase, isoform 1b, alpha2 subunit	-0.36	0.84	1.29	0.75	0.49	0.29	0.28	0.29	0.29	0.28
SET and MYND domain containing 1	-1.35	-0.99	2.00	-0.29	0.05	0.52	0.49	0.52	0.50	0.49
H3022G07	-2.07	0.78	2.49	0.18	0.99	0.90	0.87	0.91	0.90	0.87
H3018H07	0.49	1.38	2.17	1.21	1.44	0.51	0.50	0.52	0.51	0.50
RIKEN cDNA 2410025L10 gene	-0.51	1.10	1.59	0.57	0.46	0.50	0.49	0.51	0.50	0.48
H3118G03	-1.30	-0.11	0.03	-0.43	0.07	0.30	0.27	0.28	0.27	0.27
zona pellucida glycoprotein 2	-2.64	0.03	1.00	-1.01	-0.17	0.79	0.77	0.81	0.79	0.80
H3018H06	0.24	1.70	2.41	1.23	1.66	0.56	0.54	0.57	0.56	0.54

Cluster 13 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
H3083A02	2.40	2.80	2.01	1.76	1.74	0.54	0.53	0.55	0.54	0.52
U2 small nuclear RNA auxiliary factor 1-like 4	1.36	0.09	0.89	0.79	0.58	0.31	0.27	0.28	0.27	0.27
H3085A06	1.52	0.07	0.22	0.30	0.51	0.29	0.29	0.30	0.30	0.29
solute carrier family 4 (anion exchanger), member 2	1.65	0.82	0.90	-0.03	0.42	0.32	0.30	0.31	0.30	0.29
H3050C04	1.75	1.24	0.76	0.57	0.79	0.25	0.24	0.25	0.25	0.24
aldo-keto reductase family 1, member B3 (aldose reductase)	2.22	1.41	0.89	0.78	1.15	0.48	0.46	0.48	0.48	0.48
H3134D05	1.54	0.96	0.68	0.55	0.33	0.29	0.26	0.27	0.26	0.25
H3154D05	1.05	0.03	-0.24	-0.44	-0.27	0.27	0.26	0.27	0.27	0.26
H3040D12	2.93	0.57	0.25	0.27	0.12	0.39	0.38	0.40	0.39	0.38
H3048D06	1.42	1.18	0.54	0.60	0.46	0.30	0.30	0.30	0.30	0.29
H3148F08	0.44	-0.88	-0.60	-0.84	-0.75	0.32	0.32	0.33	0.32	0.32
H3134F03	1.61	1.32	0.15	0.54	0.48	0.45	0.40	0.42	0.40	0.39
golgi associated, gamma adaptin ear containing, ARF binding protein 2	2.51	2.76	2.63	2.64	2.24	0.50	0.49	0.51	0.50	0.48
H3052E11	3.06	2.75	2.07	2.38	1.88	0.52	0.51	0.53	0.52	0.51
H3132F11	1.64	1.53	0.59	0.54	0.60	0.30	0.30	0.30	0.30	0.29
H3001E12	2.86	2.40	2.14	2.09	1.88	0.53	0.51	0.53	0.53	0.51
transferrin receptor	2.96	2.48	2.80	2.15	2.54	0.39	0.38	0.40	0.39	0.38
G protein-coupled receptor 1	1.53	0.61	0.95	0.86	0.81	0.30	0.27	0.28	0.27	0.26
H3001G10	0.89	0.88	0.90	0.33	0.81	0.20	0.20	0.21	0.20	0.20
H3034G05	1.39	1.77	1.04	1.32	1.11	0.31	0.30	0.31	0.31	0.30

Cluster 14 Description	Value					Standard Error				
	D03	D07	D14	D21	D28	D03	D07	D14	D21	D28
Z3003A08	0.68	0.94	0.63	1.19	1.24	0.27	0.25	0.26	0.26	0.26
Mus musculus MAP/microtubule affinity- regulating kinase 4 (Mark4), mRNA	0.84	0.32	1.01	1.44	1.45	0.25	0.24	0.26	0.25	0.24
H3041A03	0.38	0.42	-0.09	1.21	1.36	0.40	0.39	0.41	0.40	0.39
poliovirus receptor-related 3	0.84	0.50	-0.15	0.96	0.67	0.28	0.27	0.27	0.27	0.26
H3024E09	0.98	0.67	0.67	1.30	1.71	0.36	0.35	0.37	0.36	0.35
RIKEN cDNA 2410030K01 gene	-0.06	-0.69	-0.05	0.46	0.51	0.24	0.23	0.24	0.24	0.23
H3142F10	1.26	1.55	1.15	1.72	2.01	0.26	0.25	0.26	0.26	0.25
protease, serine, 25	1.22	0.95	0.71	1.47	0.92	0.33	0.32	0.34	0.33	0.32
src homology 2 domain- containing transforming protein C1	1.71	1.76	0.68	1.64	1.36	0.35	0.34	0.35	0.35	0.34
dystrobrevin, beta	0.98	-0.54	0.38	1.03	0.63	0.39	0.39	0.40	0.42	0.38
H3049G06	0.27	-0.47	0.16	0.92	1.01	0.27	0.27	0.28	0.27	0.28

APPENDIX C DETAILED METHODS FOR MOUSE MICROVESSEL EXPERIMENT

Our analysis of gene expression by microarray is based on an analysis of variance (ANOVA) model that incorporates terms to account for known sources of variability. This approach requires an experimental design that includes repeated measures of each sample and an interconnected hybridization scheme that enables estimation of the levels of experimental sources of error (e.g. hybridization and dye variability). In our study, each time-point sample consisted of vessels from at least two microvascular constructs prepared from microvessel fragments isolated from at least 8 *tie2:GFP* mice. Each of the two (or more) constructs for each time-point was implanted into separate SCID mice. Samples were collected from the time course shown in (Figure 4.7) and hybridized according to an interconnected scheme (Figure 4.7). As indicated, two rounds of experiments were performed, with each round utilizing RNA samples generated from explanted constructs from two separate experiments. Data from each round were analyzed together by CARMA to generate expression estimates from fitting the entire data set to the linear ANOVA model.

Microvessel isolation and culture Fat microvessel fragments (MF) are isolated using a modification of previously described methods (Hoying *et al.*, 1996). Under aseptic conditions, harvested fat tissue is washed in 0.1% BSA-PBS, finely minced with scissors and digested in 2 mg/ml collagenase + 2mg/ml BSA (essentially fatty acid free) in PBS for 8 min at 37°C with vigorous shaking. Tissue debris and large vessel pieces were removed by filtering the suspension through a sterile 500 µm pore-size nylon screen. Microvessel fragments are captured by filtration of the remaining suspension on a 30 µm pore size nylon screen and recovered by vigorous flushing of the screen surface with 0.1% BSA-PBS. The type and lot number of collagenase used was pre-determined to optimize fragment yield while maintaining microvessel structure. Microvessel fragments were suspended (12,000 - 15,000 MFs/ml) in ice cold 3 mg/ml rat tail type I collagen (BD BioSciences, Bedford, MA) prepared with DMEM (1x final) and pH-neutralized with 1M NaOH. MF/collagen suspensions were plated into individual wells (0.25 ml/well) of a 48 well plate and placed in a 37°C incubator for 20 min. to polymerize the collagen.

Implantation Microvascular constructs are implanted in the subcutaneous position on the flanks of SCID mice. Each mouse will receive two implants: a microvascular construct on one side and an avascular control collagen gel on the other side. For implantation, animals were anesthetized with an intraperitoneal injection of 2.5% Avertin. Dorsal hindlimb and lower back areas were shaved, depilated and cleaned. Using blunt dissection through a small skin incision, a subcutaneous pocket was formed between the skin and the underlying muscle anterior to the pelvis. Each pocket received a construct or a control gel. The incision was then closed with 6-0 suture and the animal allowed to recover.

Histology and histochemistry Implants were removed with the underlying muscle attached, fixed in 2% paraformaldehyde/PBS and processed into paraffin. General histology was determined on deparaffinized, 5-6 μm thick sections stained with hematoxylin and eosin. Vascular elements were identified using a rodent-specific lectin, GS-1 (*Griffonia simplicifolia* I) or the human-specific endothelial cell marker UEA-1 lectin (*Ulex Europaeus* Agglutinin I). Perivascular cells were identified with a monoclonal anti- α smooth muscle actin antibody (clone 1A4; Sigma Immunochemicals) using a horse radish peroxidase (HRP) reporter system (Sigma Immunochemicals). For the lectin and immunostaining, sections were counter-stained with 1% methyl green. Vessel density was determined by counting discreet, GS-1 positive structures in the implanted construct from at least 5 different fields (of defined area) per section from two different implants. Individual counts were divided by the area of each field and averaged for each time point.

Ink perfusion Mice containing implants were anesthetized with Avertin and placed supine on a dissecting stage. The chest was opened and a catheter (PE 60 tubing) placed into the left ventricle. The mouse was perfused with PBS containing 10U/ml heparin and 10 μM sodium nitroprusside until the perfusate was clear of blood. India ink (Speedball #3398; Hunt Manufacturing Co., Statesville, NC), dialyzed against PBS and filtered through #1 Whatman paper, was perfused into the mouse at a maintained pressure of 90- 100 mm Hg until all tissues in the mouse appeared dark (usually this required 2-3 ml of ink solution). After ink perfusion, the implant was excised and fixed in 4% paraformaldehyde in PBS for 45 min. at 4°C. The fixed implants were rinsed in cold PBS, sliced longitudinally and placed in 100% glycerol for 20 min. to clarify the constructs. The two halves, cut-side up, were sandwiched between a microscope slide and coverslip for viewing with a standard light microscope.

En bloc immunohistochemistry Cultured constructs or explants were rinsed in PBS and fixed for 1 hr in 2% paraformaldehyde in PBS or 4% paraformaldehyde in PBS, respectively. Constructs were washed in cold PBS three times for 15 min. each and placed in blocking buffer (5% nonfat dry milk and 1.5% BSA in TBST buffer) overnight at 4°C. Constructs were incubated overnight at 4°C with a primary antibody directed against rat MHC (clone # OX-18) and conjugated to biotin, diluted 1:50 in blocking buffer. Following three, 1 hour washes at RT with blocking buffer, constructs were incubated with streptavidin conjugated to Oregon Green (1:400 dilution; Molecular Probes, Portland, OR) in blocking buffer for 2 hr at RT. Finally, constructs were washed twice with blocking buffer for 30 min each and twice with PBS for 30 min each, all at RT. Stained constructs were sandwiched between a microscope slide and coverslip for viewing with a standard epifluorescence microscope.

Lectin-perfusion staining. Mice were perfusion-fixed through a polyethylene catheter placed in the left ventricle. The vasculature was perfused at constant pressure (100-120 mmHg) with 20 ml of 1% paraformaldehyde and 0.5% glutaraldehyde in

phosphate-buffered saline (PBS). The animals were then perfused sequentially with each of the following; 20 ml of PBS with sodium nitroprusside (1×10^{-5} M) with heparin (10 units/ml); 20 ml of PBS and 1% BSA; 10 ml of biotinylated-lectin (20 μ g/ml final concentration; Vector Laboratories-see below) in PBS with 1% BSA; 20 ml of PBS with 1% BSA. Following the series of perfusions, the implants were dissected free from surrounding tissue and removed. The specimens were permeabilized with 0.5% Triton X-100 in PBS for 10 min and then incubated in avidin-biotin, horseradish peroxidase complex (1:100; Vectastain ABC kit, Vector Laboratories) in PBS overnight at room temperature. After washing with PBS for 1 hour, implants were then reacted for 5 to 10 minutes with diaminobenzidine substrate chromagen system (Dako Corporation) in buffer. Finally, specimens were washed in water for 15 min and then cleared in glycerol. Cleared specimens were placed on glass microscope slides and cover slipped for imaging.

Microarray hybridization Implants are harvested, minced, and submerged in RNAlater[®] (Ambion, Austin, TX). Prior to RNA isolation, tissue was removed from RNAlater[®]. RNA isolation was performed according to the protocol provided with RNA Bee[®] (Tel-Test, Inc., Friendswood, TX). Briefly, the tissue was placed in a 6 ml round-bottom tube and RNA Bee was added at 2 ml per 100 mg. Using a tissue homogenizer, tissue was ground in RNA Bee to free RNA from cells. The homogenized mixture was then added to a 2 ml Phase Lock Gel[®] (Eppendorf, Hamburg, Germany) tube to provide a stable barrier between the organic phase and the nucleic-acid aqueous phase. Chloroform was added at 200 μ l per 2 ml of RNA Bee solution. The tube was shaken for 15 seconds and centrifuged at 12,000 x g for 15 minutes at 4 °C. Upon completion of centrifugation, the aqueous layer was decanted and placed in a fresh tube. An equal volume of 100% isopropanol was added and mixed by inverting the tube several times. The tube was again placed in the centrifuge and spun for 15 minutes to pellet the RNA. The pellet was briefly washed with a 70% EtOH solution. EtOH was removed and the pellet was allowed to air dry during which time any residual EtOH evaporated. The pellets were then resuspended in nuclease-free water to a final concentration of 2-2.5 mg/ml.

RNA is amplified using the MessageAmp aRNA kit (Ambion, Austin, TX) that is based on the RNA amplification protocol developed in the laboratory of Dr. James Eberwine(7). Briefly, first strand cDNA synthesis was performed by priming 5 mg of total RNA with T7 Oligo (dT). Utilizing the T7 promoter and DNA polymerase, cDNA was converted to double-stranded DNA (dsDNA). During the cDNA purification step, Ambion's clean-up columns were replaced with MinElute[®] PCR purification columns (Qiagen, Valencia, CA) that permit the elution of cDNA in as little as 9 ml of nuclease-free H₂O. These columns were chosen because they eliminate the need to perform a vacuum centrifuge concentration. Multiple copies of aRNA were generated from the double-stranded cDNA during an overnight *in vitro* transcription using the supplied T7 NTPs and T7 enzyme mix. Following the amplification step, aRNA was purified with the provided clean-up columns and accompanying protocol. We have found that this

amplification protocol typically generates a 1000-2000 fold increase in aRNA from the original mRNA contained within the starting material.

Quantity and quality of RNA was determined by spectrophotometry. RNA was used if it had a concentration of 2 mg/ml and a 260/280 ratio of 1.6. Reverse transcription reactions were performed using EndoFree RT (Ambion) during which amino allyl dUTPs (2 mM) and dNTP's (10 mM dATP, dGTP, dCTP, and 3 mM dTTP) were incorporated. RNA was primed using random hexamers (Integrated DNA Technologies, Inc., Coralville, IA). To increase signal, reactions were run for 2 hours at 42 °C, contrary to the 48 °C recommended for the RT enzyme. After 2 hours, samples were denatured at 95°C for 5 minutes and immediately transferred to ice. Base hydrolysis of remaining RNA was performed by addition of 8.6 ml of 1M NaOH and 8.6 ml of 0.5M EDTA, pH 8, and incubated at 65 °C for 15 minutes. The solution was neutralized by adding 8.6 ml of 1M HCl. Amino allyl modified cDNA was purified using PCR purification columns (Qiagen, Valencia, CA). cDNA samples were brought up to 100 ml with Milli-Q H₂O to which 500 ml of Buffer PB was added. The PCR purification protocol was followed exactly, with the exception of substituting 75% EtOH for Buffer PE as a wash solution. cDNA was eluted off the column by using Milli-Q H₂O at a pH of 8.0.

After cDNA purification, samples were dried to 1-2 ml by vacuum centrifugation. Samples were then resuspended in 3 ml of Sodium bicarbonate (NaHCO₃)(25 mg/ml). Lyophilized Alexa dyes 546 and 647 (Molecular Probes, Eugene, OR) were diluted in 250 ml of DMSO, of which 5 ml of the appropriate dye were added to the cDNA/NaHCO₃ solution. The solution was allowed to sit in the dark for 1 hour. After 1 hour the samples were brought up to 50 ml, after which the samples labeled with 546 and 647 were combined and purified as described above, the only exception being that one extra wash with 75% EtOH was administered. To the eluate, an equal volume of 2X hybridization buffer (8X SSC, 60% Formamide, 0.2% SDS) was added along with 10 mg of Cot-1 DNA and 10 mg of poly dA. This solution was applied to a microarray slide using a GeneTac hybridization station (Genomic Solutions, Inc., Ann Arbor, MI) and hybridized at 47°C overnight. The slides were washed using two wash solutions (1X SSC, 0.1% SDS and 0.1X SSC, 0.01% SDS) by first passing the solutions over the slide for 20 seconds and then holding them on the slide for 30 seconds. Slides were rinsed in 0.1X SSC and dried.

ProLong anti-fade (Molecular Probes, Eugene, OR) was prepared per supplied protocol. To component A, 1 ml of glycerol stock solution was added and the tube was vortexed for resuspension. After vortexing, the tube was spun to eliminate any bubbles that may have formed. To a prepared array, 65ml of ProLong was added in a continuous stream down the center of the slide. A cover slip cleaned with 2M NaOH was placed over the array and the ProLong was allowed to evenly spread under the cover slip. ProLong was dried 4 hours to overnight before the slide was scanned. Hybridized, and Prolong-treated slides were scanned using a white light/CCD based scanner (Applied Precision, Inc., Issaquah, WA). Random areas of the slide were examined to determine

proper exposure for each channel. Once exposure times were established, the printed region was scanned and the data was collected. Using the spot finding analysis software (MolecularWare, Inc., Cambridge, MA), signal intensities for each spot were calculated. From that data, location specific signal intensities across the slide were generated and analyzed by the ALM.

In situ hybridization Non-radioactive *I* situ hybridization will be performed on whole mount and sectioned embryos using anti-sense and sense digoxigenin (DIG)-labeled riboprobes. Riboprobes will be *in vitro* transcribed using DIG-modified UTPs and Sp6 or T7 RNA polymerases for library clones used to manufacture the microarrays. Hybridizations on whole mount and sectioned tissues will be performed using published methods (Nieto *et al.*, 1996). Tissues are rinsed in PBS, fixed in acidic 60% EtOH, 30% formaldehyde. For tissues to be sectioned, samples are dehydrated in EtOH and embedded in Paraplast. Samples are treated with proteinase K (1mg/ml) for 7 min at 37°C and incubated with DIG-labeled riboprobes overnight at 65°C in standard hybridization buffer. Samples are washed with 50% formamide/1X SSC/.1% Tween 20 and blocked with MABT/2% blocking reagent with 2% goat serum. Positive hybridization is detected with a colorimetric enzyme reporter system targeting the DIG label.

APPENDIX D GLOSSARY OF TERMS

Additive linear model – A model containing only first order polynomial terms that are added together. These models are often used to process microarray data.

Agglomerative hierarchical clustering – A process in which genes are grouped together based on the similarity of their expression profiles. Clusters are formed by recursively combining the two genes, two clusters, or a gene and a cluster, whose expression profiles are most similar.

Analysis of variance – A statistical process that is used to determine the statistical significance of the difference between the means of two or more groups.

Angiogenesis – The formation new blood vessels from existing blood vessels.

ANOVA – See analysis of variance.

Array – See microarray.

Array term – The term in the additive linear model that accounts for variability between hybridizations.

Background subtraction – See local background subtraction.

Biological replication - Hybridizations using samples obtained from different biological specimens such as individual mice, or separate tissue culture flasks, which have received the same treatment or represent the same condition

Distance metric - Some measurement of dissimilarity, such as Euclidean distance or angle, between the m -dimensional vector that represent the m measurements for each of the n genes being compared.

Dye – See fluorochrome.

Dye term – The term in the additive linear model that accounts for variability between fluorochromes due to fluorochrome chemistry, coefficient of extinction, incorporation efficiency, photobleachability, scanner sensitivity, etc.

EST – See expressed sequence tag.

Expressed sequence tag – A sequence of transcribed DNA sequence that may or may not be protein-coding.

Fluorochrome – A molecule that when excited at one wavelength gives off light in another wavelength. The molecules are often attached to other molecules (such as amine groups) in order to bind nucleic acids.

Fluorophore – See fluorochrome.

Gene ontology – A set of structured vocabularies (ontologies) used to describe gene products in terms of their biological processes, cellular components, and molecular functions by species. Associations between these ontologies and gene products are designed to represent the existing knowledge about a gene product in each of these three areas.

Gene term – The term in the additive linear model that accounts for variability due to differences in the nucleic acid sequence between ESTs.

GO – See gene ontology.

GoMiner – A software tool that looks up the Gene Ontology (GO) information for each gene and assigns a statistical significance to each category based on the percentage of genes in the category that are considered differentially expressed.

Hierarchical clustering – See agglomerative hierarchical clustering.

Hybridization – The process of depositing a solution containing two fluorescently tagged RNA or DNA samples onto the surface of a microarray and incubating for 8 to 24 hours in order to allow sequences within the samples to bind to the complementary sequences (probes) on the microarray.

Linear model – See additive linear model.

Linlog – A type of transformation that performs a linear transformation on numbers below a specified cutoff value, and a log base 2 transformation for numbers greater than or equal to the specified cutoff value.

Local background subtraction – The process by which the median intensity of the area surrounding each spot is subtracted from the mean intensity of the area within each spot.

Lowess – Locally weighted linear regression used to normalize microarray data. This type of normalization allows the normalization factor to continuously adjust over the range of the measured intensities.

Maturation – The process through which a newly formed tube of endothelial cells with minimal mural cell coverage develops into a functional element of the vasculature (i.e. vein, venule, capillary, arteriole, artery).

Microarray – A glass microscope slide whose surface has been derivatized to bind DNA, upon which thousands of spots, each consisting of an expressed sequence tag, have been deposited.

PCA – See principal component analysis.

Principal component analysis – A statistical technique for determining the fundamental variables within a microarray dataset that best explain the differences in gene expression between the varieties (conditions).

Probe – One spot on the microarray containing millions of copies of an expressed sequence tag.

R – An open source programming language that incorporates a wide variety of statistical and graphing techniques.

Rand statistic – A measure of how well a given set of clusters matches the known clusters

in the data. The formula is: $\frac{\sum_{i=1}^n \sum_{j>i}^n \delta_{ij}}{n(n-1)/2}$ where δ_{ij} equals 1 if the genes corresponding to the row and column in the matrix were clustered together in the calculated clusters and in the assigned clusters for each of the n genes.

Remodeling – See vascular remodeling.

Sample – A small quantity of cells or tissues from which RNA is extracted and used in one or more hybridizations.

Spot – One small circular region on the microarray that is created during the printing process when a printing pin deposits a nanoliter volume of a solution containing millions of copies of an expressed sequence tag.

Technical replication – Multiple measurements of the same biological samples. There are several levels of technical replication including multiple copies of the same EST, multiple ESTs for the same transcript, and multiple hybridizations using the same biological samples.

Transcript – A sequence of nucleotides that is created when a gene is transcribed into mRNA.

Variety – The distinguishing feature of interest between samples (i.e. differences in treatment, dosage, time, genotype, etc.), which is also referred to as condition.

Variety term – The term in the additive linear model that accounts for differences due to the distinguishing feature of interest between samples (i.e. differences in treatment, dosage, time, genotype, etc.).

Vascular remodeling – The process by which a vascular network adapts to meet the metabolic demands of a tissue and the flow requirements of the network.

REFERENCES

A.H.Sturtevant (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43-59.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merrill CR, Wu A, Olde B, Moreno RF, & . (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-1656.

Alberts R, Fu J, Swertz MA, Lubbers LA, Albers CJ, & Jansen RC (2005). Combining microarrays and genetic analysis. *Brief Bioinform* **6**, 135-145.

Algire GH & Chalkley HW (1945). Vascular reactions of normal and malignant tumors in vivo. I. Vascular reactions of mice to wounds and to normal and neoplastic transplants. *J Nat Cancer Inst* **6**, 73.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, & Staudt LM (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

Alon T, Hemo I, Itin A, Pe'er J, Stone J, & Keshet E (1995). Vascular endothelial growth factor acts as a survival factor for newly formed retinal vessels and has implications for retinopathy of prematurity. *Nat Med* **1**, 1024-1028.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, & Levine AJ (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* **96**, 6745-6750.

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

Anderberg MR (1973). *Cluster Analysis for Applications* Academic Press, New York.

Applied Precision. Applied Precision ArrayWoRx Scanner.
<http://www.appliedprecision.com/lifescience/arrayworx.html> . 2005.

Arbiser JL (1996). Angiogenesis and the skin: a primer. *J Am Acad Dermatol* **34**, 486-497.

Armstrong NJ & van de Wiel MA (2004). Microarray data analysis: from hypotheses to conclusions using gene expression data. *Cell Oncol* **26**, 279-290.

Armulik A, Abramsson A, & Betsholtz C (2005). Endothelial/pericyte interactions. *Circ Res* **97**, 512-523.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, & Sherlock G (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29.

Baldi P & Long AD (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-519.

Bates D. The R Environment. 2005.

Baumgartner I, Schainfeld R, & Graziani L (2005). Management of peripheral vascular disease. *Annu Rev Med* **56**, 249-272.

Ben-Hur A, Elisseeff A, & Guyon I (2002). A stability based method for discovering structure in clustered data. *Pac Symp Biocomput* 6-17.

Benjamin LE, Golijanin D, Itin A, Pode D, & Keshet E (1999). Selective ablation of immature blood vessels in established human tumors follows vascular endothelial growth factor withdrawal. *J Clin Invest* **103**, 159-165.

Benjamin LE, Hemo I, & Keshet E (1998). A plasticity window for blood vessel remodelling is defined by pericyte coverage of the preformed endothelial network and is regulated by PDGF-B and VEGF. *Development* **125**, 1591-1598.

Benjamini Y & Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289-300.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, & Sondak V (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536-540.

Bix G & Iozzo RV (2005). Matrix revolutions: "tails" of basement-membrane components with angiostatic functions. *Trends Cell Biol* **15**, 52-60.

Bjarnegard M, Enge M, Norlin J, Gustafsdottir S, Fredriksson S, Abramsson A, Takemoto M, Gustafsson E, Fassler R, & Betsholtz C (2004). Endothelium-specific ablation of PDGFB leads to pericyte loss and glomerular, cardiac and placental abnormalities. *Development* **131**, 1847-1857.

Blashfield RK & Aldenderfer MS (1978). The Literature on Cluster Analysis. *Multivariate Behavioral Research* **13**, 271-295.

Boguski MS. Genomics Website.

http://www.elsevier.com/wps/find/journaldescription.cws_home/622838/description#description . 2005.

Botstein D, White RL, Skolnick M, & Davis RW (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314-331.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridg RB, Kirchner J, Fearon K, Mao J, & Corcoran K (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* **18**, 630-634.

Brouty-Boye D & Zetter BR (1980). Inhibition of cell motility by interferon. *Science* **208**, 516-518.

Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., & Haussler D (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97**, 262-267.

Buschmann I & Schaper W (2000). The pathophysiology of the collateral circulation (arteriogenesis). *J Pathol* **190**, 338-342.

Cao Y (2001). Endogenous angiogenesis inhibitors and their therapeutic implications. *Int J Biochem Cell Biol* **33**, 357-369.

Carmeliet P (2000). Mechanisms of angiogenesis and arteriogenesis. *Nat Med* **6**, 389-395.

Carmeliet P, Ferreira V, Breier G, Pollefeyt S, Kieckens L, Gertsenstein M, Fahrig M, Vandenhoeck A, Harpal K, Eberhardt C, Declercq C, Pawling J, Moons L, Collen D, Risau W, & Nagy A (1996). Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele. *Nature* **380**, 435-439.

Carninci P, Kasukawa T, et al. (2005). The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563.

Chen JJ, DeLongchamp RR, Tsai CA, Hsueh HM, Sistare F, Thompson KL, Desai VG, & Fuscoe JC (2004). Analysis of variance components in gene expression data. *Bioinformatics* **20**, 1436-1446.

Cheng Y & Church GM (2000). Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* **8**, 93-103.

Chiang PW, Song WJ, Wu KY, Korenberg JR, Fogel EJ, Van Keuren ML, Lashkari D, & Kurnit DM (1996). Use of a fluorescent-PCR reaction to detect genomic sequence copy number and transcriptional abundance. *Genome Res* **6**, 1013-1026.

Churchill GA (2002). Fundamentals of experimental design for cDNA microarrays. *Nat Genet* **32 Suppl**, 490-495.

Cleveland WS (1979). Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association* **74**, 829-836.

Cohen SN, Chang AC, Boyer HW, & Helling RB (1973). Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A* **70**, 3240-3244.

Cole SW, Galic Z, & Zack JA (2003). Controlling false-negative errors in microarray differential expression analysis: a PRIM approach. *Bioinformatics* **19**, 1808-1816.

Collins FS, Morgan M, & Patrinos A (2003). The Human Genome Project: lessons from large-scale biology. *Science* **300**, 286-290.

Conway EM, Collen D, & Carmeliet P (2001). Molecular mechanisms of blood vessel growth. *Cardiovasc Res* **49**, 507-521.

Cormack RM (1971). A review of classification. *Journal of Royal Statistical Society Series A*, 321-367.

Crick FH (1958). On protein synthesis. *Symp Soc Exp Biol* **12**, 138-163.

Cross MJ, Dixelius J, Matsumoto T, & Claesson-Welsh L (2003). VEGF-receptor signal transduction. *Trends Biochem Sci* **28**, 488-494.

Crum R, Szabo S, & Folkman J (1985). A new class of steroids inhibits angiogenesis in the presence of heparin or a heparin fragment. *Science* **230**, 1375-1378.

Cui X & Churchill GA. How many mice and how many arrays? Replication in mouse cDNA microarray experiments. CAMDA '02 meeting. 2002. Durham, NC.

Cui X, Hwang JT, Qiu J, Blades NJ, & Churchill GA (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59-75.

Cui X, Kerr MK, & Churchill GA (2003). Transformations for cDNA Microarray Data. *Statistical Applications in Genetics and Molecular Biology* **2**.

de Koning DJ & Haley CS (2005). Genetical genomics in humans and model organisms. *Trends Genet* **21**, 377-381.

Delmar P, Robin S, & Daudin JJ (2005). VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics* **21**, 502-508.

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, & Trent JM (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**, 457-460.

DeRisi JL, Iyer VR, & Brown PO (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686.

Dettling M & Buhlmann P (2002). Supervised clustering of genes. *Genome Biol* **3**, RESEARCH0069.

Dewey TG & Galas DJ (2001). Dynamic models of gene expression and classification. *Funct Integr Genomics* **1**, 269-278.

Ding Y & Wilkins D (2004). The effect of normalization on microarray data analysis. *DNA Cell Biol* **23**, 635-642.

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, & . (1987). A genetic linkage map of the human genome. *Cell* **51**, 319-337.

Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, & Tainsky MA (2003). Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics* **19**, 1348-1359.

Driver HE & Kroeber A.L. (1932). Quantitative Expression of Cultural Relationships. *University of California Publications in American Archaeology and Ethnology* **31**, 211-256.

Dudoit S & Fridlyand J (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* **3**, RESEARCH0036.

Duggan DJ, Bittner M, Chen Y, Meltzer P, & Trent JM (1999). Expression profiling using cDNA microarrays. *Nat Genet* **21**, 10-14.

Dvorak HF (2005). Angiogenesis: update 2005. *J Thromb Haemost* **3**, 1835-1842.

Efron B & Tibshirani R (2002). Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol* **23**, 70-86.

Eisen MB (1998). *Cluster and TreeView Manual* Stanford University.

Eisen MB, Spellman PT, Brown PO, & Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868.

Ferrara N, Carver-Moore K, Chen H, Dowd M, Lu L, O'Shea KS, Powell-Braxton L, Hillan KJ, & Moore MW (1996). Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature* **380**, 439-442.

Folkman J (1985). Tumor angiogenesis. *Adv Cancer Res* **43**, 175-203.

Folkman J, Merler E, Abernathy C, & Williams G (1971). Isolation of a tumor factor responsible for angiogenesis. *J Exp Med* **133**, 275-288.

Fong GH, Rossant J, Gertsenstein M, & Breitman ML (1995). Role of the Flt-1 receptor tyrosine kinase in regulating the assembly of vascular endothelium. *Nature* **376**, 66-70.

Form DM, Pratt BM, & Madri JA (1986). Endothelial cell proliferation during angiogenesis. In vitro modulation by basement membrane components. *Lab Invest* **55**, 521-530.

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

Garmy-Susini B, Jin H, Zhu Y, Sung RJ, Hwang R, & Varner J (2005). Integrin alpha4beta1-VCAM-1-mediated adhesion between endothelial and mural cells is required for blood vessel maturation. *J Clin Invest* **115**, 1542-1551.

Genomics Proteomics Bioinformatics (2004). Timeline of genomics (1977-2004). *Genomics Proteomics Bioinformatics* **2**, 256-267.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, & Zhang J (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80.

Gimbrone MA, Jr., Leapman SB, Cotran RS, & Folkman J (1972). Tumor dormancy in vivo by prevention of neovascularization. *J Exp Med* **136**, 261-276.

Good DJ, Polverini PJ, Rastinejad F, Le Beau MM, Lemons RS, Frazier WA, & Bouck NP (1990). A tumor suppressor-dependent inhibitor of angiogenesis is immunologically and functionally indistinguishable from a fragment of thrombospondin. *Proc Natl Acad Sci U S A* **87**, 6624-6628.

Gruionu G, Hoying JB, Pries AR, & Secomb TW (2005). Structural remodeling of mouse gracilis artery after chronic alteration in blood supply. *Am J Physiol Heart Circ Physiol* **288**, H2047-H2054.

Guigo R & Reese MG (2005). EGASP: collaboration through competition to find human genes. *Nat Methods* **2**, 575-577.

Hazen SP, Hawley RM, Davis GL, Henrissat B, & Walton JD (2003). Quantitative trait loci and comparative genomics of cereal cell wall composition. *Plant Physiol* **132**, 263-271.

He W (2004). A spline function approach for detecting differentially expressed genes in microarray data analysis. *Bioinformatics* **20**, 2954-2963.

Heid CA, Stevens J, Livak KJ, & Williams PM (1996). Real time quantitative PCR. *Genome Res* **6**, 986-994.

Heldin CH & Westermark B (1999). Mechanism of action and in vivo role of platelet-derived growth factor. *Physiol Rev* **79**, 1283-1316.

Hellstrom M, Kalen M, Lindahl P, Abramsson A, & Betsholtz C (1999). Role of PDGF-B and PDGFR-beta in recruitment of vascular smooth muscle cells and pericytes during embryonic blood vessel formation in the mouse. *Development* **126**, 3047-3055.

Heng MC, Harker J, Csathy G, Marshall C, Brazier J, Sumampong S, & Paterno GE (2000). Angiogenesis in necrotic ulcers treated with hyperbaric oxygen. *Ostomy Wound Manage* **46**, 18-2.

Hershey AD & Chase M (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**, 39-56.

Hewitson KS & Schofield CJ (2004). The HIF pathway as a therapeutic target. *Drug Discov Today* **9**, 704-711.

Heyer LJ, Kruglyak S, & Yooseph S (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* **9**, 1106-1115.

Hieter P & Boguski M (1997). Functional genomics: it's all how you read it. *Science* **278**, 601-602.

Hirschi KK & D'Amore PA (1996). Pericytes in the microvasculature. *Cardiovasc Res* **32**, 687-698.

Hoch RV & Soriano P (2003). Roles of PDGF in animal development. *Development* **130**, 4769-4784.

Horimoto K & Toh H (2001). Statistical estimation of cluster boundaries in gene expression profile data. *Bioinformatics* **17**, 1143-1151.

Hoying JB, Boswell CA, & Williams SK (1996). Angiogenic potential of microvessel fragments established in three-dimensional collagen gels. *In Vitro Cell Dev Biol Anim* **32**, 409-419.

Huber W, Von Heydebreck A, Sultmann H, Poustka A, & Vingron M (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-S104.

Hughes AL & Nei M (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167-170.

Hunter RL & Markert CL (1957). Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. *Science* **125**, 1294-1295.

Ingber DE (2002). Mechanical signaling and the cellular response to extracellular matrix in angiogenesis and cardiovascular physiology. *Circ Res* **91**, 877-887.

J.Craig Venter Institute. Genetics and Genomics Timeline.
<http://www.genomenetwork.org/resources/timeline/index.php> . 2005.

Jackson DA, Symons RH, & Berg P (1972). Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proc Natl Acad Sci U S A* **69**, 2904-2909.

JACOB F & Monod J (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-356.

Jain RK (2005). Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. *Science* **307**, 58-62.

Jang YC, Arumugam S, Ferguson M, Gibran NS, & Isik FF (1998). Changes in matrix composition during the growth and regression of human hemangiomas. *J Surg Res* **80**, 9-15.

Jeffreys AJ, Wilson V, & Thein SL (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**, 67-73.

Kargul GJ, Dudekula DB, Qian Y, Lim MK, Jaradat SA, Tanaka TS, Carter MG, & Ko MS (2001). Verification and initial annotation of the NIA mouse 15K cDNA clone set. *Nat Genet* **28**, 17-18.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engstrom PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, & Wahlestedt C (2005). Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566.

Keller EF (2000). *The century of the gene* Harvard University Press, Cambridge, MA.

Kendzioriski CM, Newton MA, Lan H, & Gould MN (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med* **22**, 3899-3914.

Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ, & Churchill GA (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* **12**, 203-217.

Kerr MK & Churchill GA (2001a). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad Sci U S A* **98**, 8961-8965.

Kerr MK & Churchill GA (2001b). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183-201.

Kerr MK, Martin M, & Churchill GA (2000). Analysis of variance for gene expression microarray data. *J Comput Biol* **7**, 819-837.

Kimura H, Weisz A, Kurashima Y, Hashimoto K, Ogura T, D'Acquisto F, Addeo R, Makuuchi M, & Esumi H (2000). Hypoxia response element of the human vascular endothelial growth factor gene mediates transcriptional regulation by nitric oxide: control of hypoxia-inducible factor-1 activity by nitric oxide. *Blood* **95**, 189-197.

Kluger Y, Basri R, Chang JT, & Gerstein M (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res* **13**, 703-716.

Kubota S, Kawata K, Yanagita T, Doi H, Kitoh T, & Takigawa M (2004). Abundant retention and release of connective tissue growth factor (CTGF/CCN2) by platelets. *J Biochem (Tokyo)* **136**, 279-282.

Kuiper KF & Fisher L (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics* **31**, 777-783.

Landegren U, Nilsson M, & Kwok PY (1998). Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Res* **8**, 769-776.

Landry JR, Kinston S, Knezevic K, Donaldson IJ, Green AR, & Gottgens B (2005). Fli1, Elf1 and Ets1 regulate the proximal promoter of the LMO2 gene in endothelial cells. *Blood*.

Lee ML, Kuo FC, Whitmore GA, & Sklar J (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A* **97**, 9834-9839.

Lee Y & Lee CK (2003). Classification of multiple cancer types by multiclass support vector machines using gene expression data. *Bioinformatics* **19**, 1132-1139.

Leveen P, Pekny M, Gebre-Medhin S, Swolin B, Larsson E, & Betsholtz C (1994). Mice deficient for PDGF B show renal, cardiovascular, and hematological abnormalities. *Genes Dev* **8**, 1875-1887.

Levine E & Domany E (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Comput* **13**, 2573-2593.

Li H, Wood CL, Getchell TV, Getchell ML, & Stromberg AJ (2004). Analysis of oligonucleotide array experiments with repeated measures using mixed models. *BMC Bioinformatics* **5**, 209.

Lindley DV (1962). Discussion on Professor Stein's Paper. *J R Stat Soc Ser B* **24**, 265-296.

Lipshutz RJ, Fodor SP, Gingeras TR, & Lockhart DJ (1999). High density synthetic oligonucleotide arrays. *Nat Genet* **21**, 20-24.

Lonnstedt I & Speed T (2002). Replicated Microarray Data. *Statistica Sinica* **12**, 31-46.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, & Frazer KA (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136-140.

Luttun A, Tjwa M, Moons L, Wu Y, ngelillo-Scherrer A, Liao F, Nagy JA, Hooper A, Priller J, De KB, Compennolle V, Daci E, Bohlen P, Dewerchin M, Herbert JM, Fava R, Matthys P, Carmeliet G, Collen D, Dvorak HF, Hicklin DJ, & Carmeliet P (2002). Revascularization of ischemic tissues by PIGF treatment, and inhibition of tumor angiogenesis, arthritis and atherosclerosis by anti-Flt1. *Nat Med* **8**, 831-840.

Maeshima Y, Colorado PC, Torre A, Holthaus KA, Grunkemeyer JA, Ericksen MB, Hopfer H, Xiao Y, Stillman IE, & Kalluri R (2000). Distinct antitumor properties of a type IV collagen domain derived from basement membrane. *J Biol Chem* **275**, 21340-21348.

Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H, Yang G, Kennedy GC, Webster TA, Cawley S, Walsh PS, Jones KW, Fodor SP, & Mei R (2004). Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1**, 109-111.

Maxam AM & Gilbert W (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560-564.

McKusick VA & Ruddle FH (1987). A new discipline, a new name, a new journal. *Genomics* **1**, 1-2.

McQuitty LL (1957). Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. *Educational and Psychological Measurement* **17**, 207-229.

McReynolds MR, Taylor-Garcia KM, Greer KA, Hoying JB, & Brooks HL (2005). Renal medullary gene expression in aquaporin-1 null mice. *Am J Physiol Renal Physiol* **288**, F315-F321.

Mendel G (1866). Versuche uber Pflanzenhybriden. *Naturforsch-Verhandlungen Brunn* **4**, 1-47.

Messier W & Stewart CB (1997). Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151-154.

Millauer B, Wizigmann-Voos S, Schnurch H, Martinez R, Moller NP, Risau W, & Ullrich A (1993). High affinity VEGF binding and developmental expression suggest Flk-1 as a major regulator of vasculogenesis and angiogenesis. *Cell* **72**, 835-846.

Miller W, Makova KD, Nekrutenko A, & Hardison RC (2004). Comparative genomics. *Annu Rev Genomics Hum Genet* **5**, 15-56.

Milligan GW (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* **45**, 325-342.

Modlich O, Prisack HB, Munnes M, Audretsch W, & Bojar H (2005). Predictors of primary breast cancers responsiveness to preoperative Epirubicin/Cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures. *J Transl Med* **3**, 32.

Molecular Devices. Molecular Devices GenePix Scanner. http://www.moleculardevices.com/pages/instruments/microarray_main.html . 2005.

Mongiati M, Sweeney SM, San Antonio JD, Fu J, & Iozzo RV (2003). Endorepellin, a novel inhibitor of angiogenesis derived from the C terminus of perlecan. *J Biol Chem* **278**, 4238-4249.

Morgan TH, Sturtevant AH, MHJ, & aBCB (1915). *The Mechanism of Mendelian Heredity* Henry Holt, New York.

Motoike T, Loughna S, Perens E, Roman BL, Liao W, Chau TC, Richardson CD, Kawate T, Kuno J, Weinstein BM, Stainier DY, & Sato TN (2000). Universal GFP reporter for the study of vascular development. *Genesis* **28**, 75-81.

Mukherjee S, Roberts SJ, & van der Laan MJ (2005). Data-adaptive test statistics for microarray data. *Bioinformatics* **21 Suppl 2**, ii108-ii114.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, & Erlich H (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* **51 Pt 1**, 263-273.

Murasawa S & Asahara T (2005). Endothelial progenitor cells for vasculogenesis. *Physiology (Bethesda)* **20**, 36-42.

Murray CD (1926). The Physiological Principle of Minimum Work. I. The Vascular System and the Cost of Blood Volume. *PNAS* **12**, 207-214.

Nature Publishing Group.

<http://www.nature.com/focus/angiogenesis/classics/antifactors.html> . 2005.

Nieto MA, Patel K, & Wilkinson DG (1996). In situ hybridization analysis of chick embryos in whole mount and tissue sections. *Methods Cell Biol* **51**, 219-235.

Nirenberg MW, Matthaei JH, & Jones OW (1962). An intermediate in the biosynthesis of polyphenylalanine directed by synthetic template RNA. *Proc Natl Acad Sci U S A* **48**, 104-109.

Notterman DA, Alon U, Sierk AJ, & Levine AJ (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res* **61**, 3124-3130.

Nyberg P, Xie L, & Kalluri R (2005). Endogenous inhibitors of angiogenesis. *Cancer Res* **65**, 3967-3979.

O'Neil D. Mendel's Genetics. http://anthro.palomar.edu/mendel/mendel_1.htm . 2005.

O'Reilly MS, Boehm T, Shing Y, Fukai N, Vasios G, Lane WS, Flynn E, Birkhead JR, Olsen BR, & Folkman J (1997). Endostatin: an endogenous inhibitor of angiogenesis and tumor growth. *Cell* **88**, 277-285.

Ovcharenko I, Nobrega MA, Loots GG, & Stubbs L (2004). ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* **32**, W280-W286.

Packard Bioscience. Packard Bioscience ScanArray 4000. <http://www.gmi-inc.com/BioTechLab/GSI%20Lumonics%20ScanArray%204000.html> . 2005.

Palijan A, Lambert R, Dutil J, Sivo Z, & Deng AY (2003). Comprehensive congenic coverage revealing multiple blood pressure quantitative trait loci on Dahl rat chromosome 10. *Hypertension* **42**, 515-522.

Pan W, Lin J, & Le CT (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol* **3**, research0022.

Parmigiani G, Garrett ES, Irizarry RA, & Zeger SL (2003). *The Analysis of Gene Expression Data: Methods and Software* Springer.

Patan S, Alvarez MJ, Schittny JC, & Burri PH (1992). Intussusceptive microvascular growth: a common alternative to capillary sprouting. *Arch Histol Cytol* **55 Suppl**, 65-75.

Patan S, Munn LL, Tanda S, Roberge S, Jain RK, & Jones RC (2001). Vascular morphogenesis and remodeling in a model of tissue repair: blood vessel formation and growth in the ovarian pedicle after ovariectomy. *Circ Res* **89**, 723-731.

Pavlidis P, Li Q, & Noble WS (2003). The effect of replication on gene expression microarray experiments. *Bioinformatics* **19**, 1620-1627.

Peirce SM & Skalak TC (2003). Microvascular remodeling: a complex continuum spanning angiogenesis to arteriogenesis. *Microcirculation* **10**, 99-111.

Peirce SM, Van Gieson EJ, & Skalak TC (2004). Multicellular simulation predicts microvascular patterning and in silico tissue assembly. *FASEB J* **18**, 731-733.

Pennacchio LA (2003). Insights from human/mouse genome comparisons. *Mamm Genome* **14**, 429-436.

Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM, & Rubin EM (2001). An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**, 169-173.

Pepper MS (1997). Manipulating angiogenesis. From basic science to the bedside. *Arterioscler Thromb Vasc Biol* **17**, 605-619.

Pepper MS, Ferrara N, Orci L, & Montesano R (1992). Potent synergism between vascular endothelial growth factor and basic fibroblast growth factor in the induction of angiogenesis in vitro. *Biochem Biophys Res Commun* **189**, 824-831.

Price RJ & Skalak TC (1996). Chronic alpha 1-adrenergic blockade stimulates terminal and arcade arteriolar development. *Am J Physiol* **271**, H752-H759.

Pries AR & Secomb TW (2005). Control of blood vessel structure: insights from theoretical models. *Am J Physiol Heart Circ Physiol* **288**, H1010-H1015.

Pries AR, Secomb TW, Gaehtgens P, & Gross JF (1990). Blood flow in microvascular networks. Experiments and simulation. *Circ Res* **67**, 826-834.

Pries AR, Secomb TW, Gessner T, Sperandio MB, Gross JF, & Gaehtgens P (1994). Resistance to blood flow in microvessels in vivo. *Circ Res* **75**, 904-915.

Pritchard CC, Hsu L, Delrow J, & Nelson PS (2001). Project normal: defining normal variance in mouse gene expression. *Proc Natl Acad Sci U S A* **98**, 13266-13271.

Qin L, Rueda L, Ali A, & Ngom A (2005). Spot detection and image segmentation in DNA microarray data. *Appl Bioinformatics* **4**, 1-11.

- Qin LX & Kerr KF (2004). Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res* **32**, 5471-5479.
- Quackenbush J (2002). Microarray data normalization and transformation. *Nat Genet* **32 Suppl**, 496-501.
- Rafii S, Meeus S, Dias S, Hattori K, Heissig B, Shmelkov S, Rafii D, & Lyden D (2002). Contribution of marrow-derived progenitors to vascular and cardiac regeneration. *Semin Cell Dev Biol* **13**, 61-67.
- Rahmenfuhrer J (2005). Image analysis for cDNA microarrays. *Methods Inf Med* **44**, 405-407.
- Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, & Golub TR (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* **98**, 15149-15154.
- Rand WM (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846-850.
- Rastinejad F, Polverini PJ, & Bouck NP (1989). Regulation of the activity of a new inhibitor of angiogenesis by a cancer suppressor gene. *Cell* **56**, 345-355.
- Raychaudhuri S, Stuart JM, & Altman RB (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput* 455-466.
- Renwick JH & Lawler SD (1955). Genetical linkage between the ABO and nail-patella loci. *Ann Hum Genet* **19**, 312-331.
- Ribatti D, Crivellato E, Roccaro AM, Ria R, & Vacca A (2004). Mast cell contribution to angiogenesis related to tumour progression. *Clin Exp Allergy* **34**, 1660-1664.
- Rocke DM & Durbin B (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* **19**, 966-972.

Rocke DM & Durbin B (2001). A model for measurement error for gene expression arrays. *J Comput Biol* **8**, 557-569.

Rohlf FJ (1974). Methods of Comparing Classifications. *Annual Review of Ecology and Systematics* **5**, 101-113.

Rosinberg A, Khan TA, Sellke FW, & Laham RJ (2004). Therapeutic angiogenesis for myocardial ischemia. *Expert Rev Cardiovasc Ther* **2**, 271-283.

Rundhaug JE (2005). Matrix metalloproteinases and angiogenesis. *J Cell Mol Med* **9**, 267-285.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, & Erlich HA (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487-491.

Sanger F, Nicklen S, & Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467.

Sax K (1923). THE ASSOCIATION OF SIZE DIFFERENCES WITH SEED-COAT PATTERN AND PIGMENTATION IN PHASEOLUS VULGARIS. *Genetics* **8**, 552-560.

Schena M, Shalon D, Davis RW, & Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.

Schena M, Shalon D, Heller R, Chai A, Brown PO, & Davis RW (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A* **93**, 10614-10619.

Scholz D, Cai WJ, & Schaper W (2001). Arteriogenesis, a new concept of vascular adaptation in occlusive disease. *Angiogenesis* **4**, 247-257.

Schultz GS & Grant MB (1991). Neovascular growth factors. *Eye* **5 (Pt 2)**, 170-180.

- Senger DR, Galli SJ, Dvorak AM, Perruzzi CA, Harvey VS, & Dvorak HF (1983). Tumor cells secrete a vascular permeability factor that promotes accumulation of ascites fluid. *Science* **219**, 983-985.
- Serini G, Ambrosi D, Giraudo E, Gamba A, Preziosi L, & Bussolino F (2003). Modeling the early stages of vascular network assembly. *EMBO J* **22**, 1771-1779.
- Shalaby F, Rossant J, Yamaguchi TP, Gertsenstein M, Wu XF, Breitman ML, & Schuh AC (1995). Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice. *Nature* **376**, 62-66.
- Sharov AA, Dudekula DB, & Ko MS (2005). A web-based tool for principal component and significance analysis of microarray data. *Bioinformatics* **21**, 2548-2549.
- Sheng Q, Moreau Y, & De MB (2003). Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19 Suppl 2**, II196-II205.
- Shepherd BR, Chen HY, Smith CM, Gruionu G, Williams SK, & Hoying JB (2004). Rapid perfusion and network remodeling in a microvascular construct after implantation. *Arterioscler Thromb Vasc Biol* **24**, 898-904.
- Sherlock G (2000). Analysis of large-scale gene expression data. *Curr Opin Immunol* **12**, 201-205.
- Smith HO & Wilcox KW (1970). A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. *J Mol Biol* **51**, 379-391.
- Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, & Hood LE (1986). Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679.
- Smolkin M & Ghosh D (2003). Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics* **4**, 36.

Smyth GK (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 3.

Smyth GK, Michaud J, & Scott HS (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067-2075.

Smyth GK & Speed T (2003). Normalization of cDNA microarray data. *Methods* **31**, 265-273.

Sokal RR & Sneath PHA (1963). *Principles of Numerical Taxonomy* W. H. Freeman, San Francisco.

Sokal RR & Michener CD (1958). A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull* **38**, 1409-1438.

Soriano P (1994). Abnormal kidney development and hematological disorders in PDGF beta-receptor mutant mice. *Genes Dev* **8**, 1888-1896.

Soslau G, Morgan DA, Jaffe JS, Brodsky I, & Wang Y (1997). Cytokine mRNA expression in human platelets and a megakaryocytic cell line and cytokine modulation of platelet function. *Cytokine* **9**, 405-411.

Sun S, Wheeler MF, Obeyesekere M, & Patrick CW, Jr. (2005). A deterministic model of growth factor-induced angiogenesis. *Bull Math Biol* **67**, 313-337.

Swinscoe JC & Carlson EC (1992). Capillary endothelial cells secrete a heparin-binding mitogen for pericytes. *J Cell Sci* **103 (Pt 2)**, 453-461.

Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, Doi H, Wood WH, III, Becker KG, & Ko MS (2000). Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc Natl Acad Sci U S A* **97**, 9127-9132.

- Tanay A, Sharan R, Kupiec M, & Shamir R (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* **101**, 2981-2986.
- Tanay A, Sharan R, & Shamir R (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18 Suppl 1**, S136-S144.
- Tanksley SD, Young ND, Paterson AH, & Bonierbale MW (1989). RFLP Mapping in Plant Breeding: New Tools for an Old Science. *Nat Biotech* **7**, 257-264.
- Taylor S & Folkman J (1982). Protamine is an inhibitor of angiogenesis. *Nature* **297**, 307-312.
- Thurston G, Baluk P, & McDonald DM (2000). Determinants of endothelial cell phenotype in venules. *Microcirculation* **7**, 67-80.
- Thurston G, Wang Q, Baffert F, Rudge J, Papadopoulos N, Jean-Guillaume D, Wiegand S, Yancopoulos GD, & McDonald DM (2005). Angiopoietin 1 causes vessel enlargement, without angiogenic sprouting, during a critical developmental period. *Development* **132**, 3317-3326.
- Toronen P, Kolehmainen M, Wong G, & Castren E (1999). Analysis of gene expression data using self-organizing maps. *FEBS Lett* **451**, 142-146.
- Tusher VG, Tibshirani R, & Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121.
- Twigger SN, Pasko D, Nie J, Shimoyama M, Bromberg S, Campbell D, Chen J, Dela CN, Fan C, Foote C, Harris G, Hickmann B, Ji Y, Jin W, Li D, Mathis J, Nenasheva N, Nigam R, Petri V, Reilly D, Ruotti V, Schauburger E, Seiler K, Slyper R, Smith J, Wang W, Wu W, Zhao L, Zuniga-Meyer A, Tonellato PJ, Kwitek AE, & Jacob HJ (2005). Tools and Strategies for Physiological Genomics - The Rat Genome Database. *Physiol Genomics*.
- U.S.Department of Health and Human Services. Health, United States, 2004. <http://www.cdc.gov/nchs/hus.htm> . 2005.

Van den Steen PE, Proost P, Wuyts A, Van Damme J, & Opdenakker G (2000). Neutrophil gelatinase B potentiates interleukin-8 tenfold by aminoterminal processing, whereas it degrades CTAP-III, PF-4, and GRO-alpha and leaves RANTES and MCP-2 intact. *Blood* **96**, 2673-2681.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, & Friend SH (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536.

Velculescu VE, Zhang L, Vogelstein B, & Kinzler KW (1995). Serial analysis of gene expression. *Science* **270**, 484-487.

Walsh DA & Pearson CI (2001). Angiogenesis in the pathogenesis of inflammatory joint and lung diseases. *Arthritis Res* **3**, 147-153.

Wang A & Gehan EA (2005). Gene selection for microarray data analysis using principal component analysis. *Stat Med* **24**, 2069-2087.

Wang S & Ethier S (2004). A generalized likelihood ratio test to identify differentially expressed genes from microarray data. *Bioinformatics* **20**, 100-104.

Wardrop SL & Brown MA (2005). Identification of two evolutionarily conserved and functional regulatory elements in intron 2 of the human BRCA1 gene. *Genomics* **86**, 316-328.

Watanabe M, Jafri A, & Fisher SA (2001). Apoptosis is required for the proper formation of the ventriculo-arterial connections. *Dev Biol* **240**, 274-288.

Waterston RH, Lindblad-Toh K, et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.

Watson JD & Crick FH (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738.

Weber JL & May PE (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* **44**, 388-396.

Wegewitz U, Gohring I, & Spranger J (2005). Novel approaches in the treatment of angiogenic eye disease. *Curr Pharm Des* **11**, 2311-2330.

Wei L, Keogh CL, Whitaker VR, Theus MH, & Yu SP (2005). Angiogenesis and stem cell transplantation as potential treatments of cerebral ischemic stroke. *Pathophysiology* **12**, 47-62.

Wilson DL, Buckley MJ, Helliwell CA, & Wilson IW (2003). New normalization methods for cDNA microarray data. *Bioinformatics* **19**, 1325-1332.

Winkler F, Kozin SV, Tong RT, Chae SS, Booth MF, Garkavtsev I, Xu L, Hicklin DJ, Fukumura D, di TE, Munn LL, & Jain RK (2004). Kinetics of vascular normalization by VEGFR2 blockade governs brain tumor response to radiation: role of oxygenation, angiopoietin-1, and matrix metalloproteinases. *Cancer Cell* **6**, 553-563.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, & Paules RS (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* **8**, 625-637.

Wright GW & Simon RM (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448-2455.

Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, & Quackenbush J (2002a). Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol* **3**, research0062.

Yang MCK, Ruan QG, Yang JJ, Eckenrode S, Wu S, McIndoe RA, & She JX (2001a). A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol Genomics* **7**, 45-53.

Yang YH, Buckley MJ, & Speed TP (2001b). Analysis of cDNA microarray images. *Brief Bioinform* **2**, 341-349.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, & Speed TP (2002b). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15.

Yang YH & Speed T (2002). Design issues for cDNA microarray experiments. *Nat Rev Genet* **3**, 579-588.

Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, & Weinstein JN (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* **4**, R28.

Zetter BR (1988). Angiogenesis. State of the art. *Chest* **93**, 159S-166S.

Zhang K & Zhao H (2000). Assessing reliability of gene clusters from gene expression data. *Functional & Integrative Genomics* **1**, 156-173.

Zhao Y & Pan W (2003). Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics* **19**, 1046-1054.

Zimdahl H, Kreitler T, Gosele C, Ganten D, & Hubner N (2002). Conserved synteny in rat and mouse for a blood pressure QTL on human chromosome 17. *Hypertension* **39**, 1050-1052.