

DESIGN AND APPLICATION OF THE KENTUCKY MICROARRAY ANALYSIS

SUITE

by

Vijay Anand Raghavan

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

In

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

November 2006

© COPYRIGHT

Vijay Anand Raghavan

2006

All Rights Reserved

APPROVAL

of a thesis submitted by

Vijay Anand Raghavan

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Division of Graduate Education.

Dr. Brendan Mumey

Approved for the Department of Computer Science

Dr. Michael Oudshoorn

Approved for the Division of Graduate Education

Dr. Carl A. Fox

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library.

If I have indicated my intention to copyright this thesis by including a copyright notice page, copying is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for permission for extended quotation from or reproduction of this thesis in whole or in parts may be granted only by the copyright holder.

Vijay Anand Raghavan

November 2006

ACKNOWLEDGMENTS

I would like to thank Dr. Brendan Mumey, my advisor for his guidance and assistance throughout the project. A special thanks to Dr. Tim Ford for his support all these years. I am grateful to my committee members Dr. Rafal Angryk and Dr. Denbigh Starkey for their support and guidance. I am indebted to Dr. Jim Lund for his time and patience, and most of all giving me freedom to explore new ideas, without him this project could have never been possible.

TABLE OF CONTENTS

1. INTRODUCTION	
Overview.....	1
Microarrays.....	2
Types of Microarrays.....	2
Spotted Microarrays.....	2
Oligonucleotide Microarrays.....	2
Manufacture of Affymetrix GENECHIP® Arrays.....	3
Overview of the Results.....	4
2. BACKGROUND	
Research on Existing Products.....	5
S-Plus® with S+ ArrayAnalyzer®.....	5
GeneChip® Operating Software (GCOS).....	6
L2L Microarray Analysis Tool.....	6
GeneSpring®.....	6
Results of the Survey.....	7
Motivation for the Kentucky Microarray Analysis Suite.....	7
Microarray Data Storage.....	7
Microarray Data in MADUK.....	8
3. DESIGN AND IMPLEMENTATION	
Array Quality Analyzer.....	9
Differentially Expressed Gene Finder.....	10
Normalization.....	11
Building a Linear Model.....	11
Calculating differences in gene expression.....	12
Identification of the differentially expressed genes.....	12
Cluster Pre-Processor.....	13
Gene Ontology Based Group Predictor and Heatmap Creator.....	13
Kentucky Microarray Analysis Suite vs other Microarray Analysis Software.....	14
4. RESULTS	
<i>Caenorhabditis elegans</i>	15
Reproduction in <i>C. elegans</i>	15
Life cycle of <i>Caenorhabditis elegans</i>	16
Evolutionary Relationship.....	16
Genome of <i>Caenorhabditis elegans</i>	17
Mitochondria.....	17
Mitochondria and Aging.....	17

TABLE OF CONTENTS – CONTINUED

Mitochondrial Genes.....	18
Aging.....	18
Effects of Aging in Humans in <i>C. elegans</i>	19
Aging and Mitochondrial Genes.....	19
Microarray Experimental Design.....	20
Analysis of Differentially Expressed Genes	21
Classification of the Differentially Expressed Genes	21
Clustering of Differentially Expressed Genes	22
Gene Ontology Analysis of the Gene Expression Data	22
Mitochondrial Gene Expression	28
5. CONCLUSIONS	
Future Work	31
REFERENCES CITED.....	32

LIST OF TABLES

Table	Page
1. Microarray expression analysis of mitochondrial genes.....	21
2. Classification of differentially expressed genes based on function.....	22

LIST OF FIGURES

Figure	Page
1. MVA Plot of the Controls.....	10
2. Clustergram of pooled differentially expressed genes from the six mitochondrial gene RNAi.....	24
3. GO Molecular Function heat map.....	25
4. GO Cellular Component heat map.....	26
5. GO Biological Process heatmap.....	27
6. Clustogram of 284 mitochondrial genes.....	29

ABSTRACT

In recent years, microarrays have become the most widely used standard in the study of gene expression. The biggest problem in microarray data analysis is the dimensionality of the data, compared to other more traditional biomedical research methods. The inherent nature of the data, and the problems associated with the microarray data analysis, has led to the development of many methods for microarray data analysis. Microarray data analysis methods are commonly classified into Class Discovery methods e.g. clustering, Class Comparison methods e.g. predicting differentially expressed genes, and Class Prediction methods e.g. classification. In this thesis, a new microarray analysis tool called Kentucky Microarray Analysis Suite that has all the three major microarray analysis methods is introduced. As a proof of concept Affymetrix array data related to aging in *C. elegans* is analyzed with the Kentucky Microarray Analysis Suite and the results are presented.

CHAPTER ONE

INTRODUCTION

Overview

In this thesis the design, development and application of the Kentucky Microarray Analysis Suite is described. The biggest problem in microarray data analysis is the dimensionality of the data, compared to the data from other more traditional biomedical research methods. The traditional machine learning based algorithms do not fare well in microarray data analysis, due to the fact that there are a small number of samples compared to the huge number of features (i.e. genes).

The inherent nature of the microarray data, and the problems associated with the microarray data analysis, led to the development of many methods for microarray data analysis. Microarray data analysis methods are commonly classified into Class Discovery methods, e.g. clustering, Class Comparison methods [1], e.g. predicting differentially expressed genes, and Class Prediction methods [2], e.g. classification.

In many microarray data analysis, these three methods are used in isolation. The aim of this project, is to develop a suite of programs to combine the above three classes of microarray data analysis to analyze Affymetrix microarray data related to aging in *C. elegans*, and to introduced a new Class Prediction method based on the Gene Ontology database.

Microarrays

Microarrays evolved from Southern blotting and in recent times microarrays have become the defacto standard to study gene expression. In a DNA microarray thousands of microscopic DNA spots known as probes are attached usually to a silicon chip, plastic, glass, or a quartz surface to profile the expression patterns of thousands of genes at the same time.

Types of Microarrays

Spotted Microarrays

Microarrays whose probes are oligonucleotides, cDNA or PCR product fragments corresponding to mRNAs are called spotted microarrays or two-channel microarrays. In Spotted microarrays the array is hybridized with cDNA from two samples that studied labeled with two different colors. The microarray is later scanned to find the genes that are up and down regulated. Spotted microarrays are usually cheaper but the main drawback of this technology is that absolute levels of gene expression cannot be observed.

Oligonucleotide Microarrays

In the case of oligonucleotide microarrays, also known as single-channel microarrays, the probes are usually engineered with sequences of known or predicted

mRNAs. Affymetrix GeneChip® Arrays are one of the better known oligonucleotide microarrays on the market. These arrays provide the absolute levels of gene expression.

Manufacture of Affymetrix GeneChip® Arrays

The GeneChip® Arrays are fabricated using a photolithographic process. The process begins with coating the quartz wafer with a chemical compound that prevents coupling between the wafer and the first nucleotide of the DNA probe being created. Light is either blocked or transmitted into specific locations of the wafer surface using lithographic masks. The surface is then treated with a solution containing adenine, thymine, cytosine, or guanine. In the regions that are deprotected through illumination, coupling occurs.

The coupled nucleotide also bears a light-sensitive protecting group, so the cycle can be repeated. In this way, the microarray is built as the probes are synthesized through repeated cycles of deprotection and coupling. The process is repeated until the probes reach their full length, usually 25 nucleotides. Commercially available arrays are typically manufactured at a density of over 1.3 million unique features per array. Depending on the demands of the experiment and the number of probes required per array, each quartz wafer can be diced into tens or hundreds of individual arrays [18].

Overview of the Results

The results obtained at the end of analysis, of the Affymetrix microarray data related to aging in *C.elegans* using Class Discovery, Class Comparison, and Class Prediction methods led to the discovery of 465 differentially expressed genes, distinct clustering of samples based on gene expression levels and Gene Ontology groups that had a significant change in their expression pattern.

CHAPTER TWO

BACKGROUND

A survey of existing Microarray data analysis tools, led to the conclusion that none of the existing tools and products meet the needs specific to the project. This led to the creation of Kentucky Microarray Analysis Suite (KMAS).

Research on Existing Products

There are many products in the market today, both commercial and open source, for microarray data analysis. A survey of the products available in the market today was done to evaluate their suitability for this project. Each product surveyed is discussed below.

S-Plus® with S+ ArrayAnalyzer®

S-Plus® with S+ ArrayAnalyzer® is a commercial product from Insightful Corporation. The product comes with programs called analytics to analyze microarray data. There are different programs for clustering and identifying the differentially expressed genes. The main problem with the product is that it has no classification and annotation methods. Since the product is commercial the license is expensive. The proprietary nature of the product makes it impossible to use the product in conjunction with the open source Microarray Database at the University of Kentucky (MADUK).

GeneChip® Operating Software (GCOS)

GeneChip® Operating Software is a commercial product from Affymetrix. The product was first developed for the Affymetrix array station. Later, many microarray data analysis features were added to the product. The proprietary nature of the product and inconsistent changes in the data format makes it impossible to use the product in conjunction with the Microarray Database at the University of Kentucky (MADUK).

L2L Microarray Analysis Tool

L2L is a free web-based tool from the University of Washington, for microarray data analysis. It is primarily a database of published microarray gene expression data. When a user enters the microarray data, a comparison is done with the existing expression data and the matched results are reported. It lacks the ability for customization and lacks methods for clustering and identification of differentially expressed genes.

GeneSpring®

GeneSpring® from Agilent is considered the gold standard in the area of microarray data analysis. It has methods for identifying differentially expressed genes and clustering. It lacks tools for annotation and classification, and the nature of the licensing agreement of the product makes it impossible to use the product with the Microarray Database at the University of Kentucky (MADUK).

Results of the Survey

Most of the tools surveyed did not have all the three methods of microarray data analysis, i.e. Class Discovery, Class Comparison and Class Prediction, in them and some of the products that had all the three methods did not integrate well with the Kentucky Microarray Database due to their proprietary nature and inherent design.

Motivation for the Kentucky Microarray Analysis Suite

The microarray analysis products mentioned before lack the ability to integrate with an existing microarray database. The better performing, feature rich products like GeneSpring are commercial with a huge licensing fee. This leaves a huge void in the area of microarray database analysis for a freely available, feature rich microarray analysis tool that is also easy to integrate with an existing microarray databases. This led to the development of KMAS is a suite of programs, written in a mixture of R and PERL for microarray data analysis. KMAS has different methods for identifying differentially expressed genes, annotation, Gene Ontology (GO) based analysis and preparing data for clustering. KMAS was born out of the idea of analyzing microarray data using all the three major class methods for microarray data analysis.

Microarray Data Storage

Microarray experiments produce a large amount of data. Storage, retrieval and tracking of microarray data present their own complexities. Over the years many

products have been released in the market for storing and retrieving microarray data. Many of the products are open source but some of them are proprietary. The Stanford Microarray database is one of the important open source products in the market. The Microarray Database at the University of Kentucky (MADUK) first started as a local implementation of the Stanford Microarray Database to store and analyze its microarray data. Later many new features were added to the original Stanford Microarray Database template. Data from MADUK can be published to GEO in MIAME-compliant MAGE-ML format.

Microarray Data in MADUK

MADUK currently accepts data from Agilent, GenePix or ScanAlyze, SpotReader, NimbleGen, and Affymetrix. In the case of Affymetrix arrays the GeneChip print, experiment file, image file, oligo intensity file, and probeset values file are needed, to enter an experiment into MADUK. MADUK also has several data analysis functions, some of them are:

1. Clustering.
2. Plotting the location of genes along the chromosome (Caryoscope).
3. Ontology Term Finder.
4. Synthetic genes etc.

CHAPTER THREE

DESIGN AND IMPLEMENTATION

The Kentucky Microarray Analysis Suite is composed of four main programs, written in a combination of PERL and R listed below.

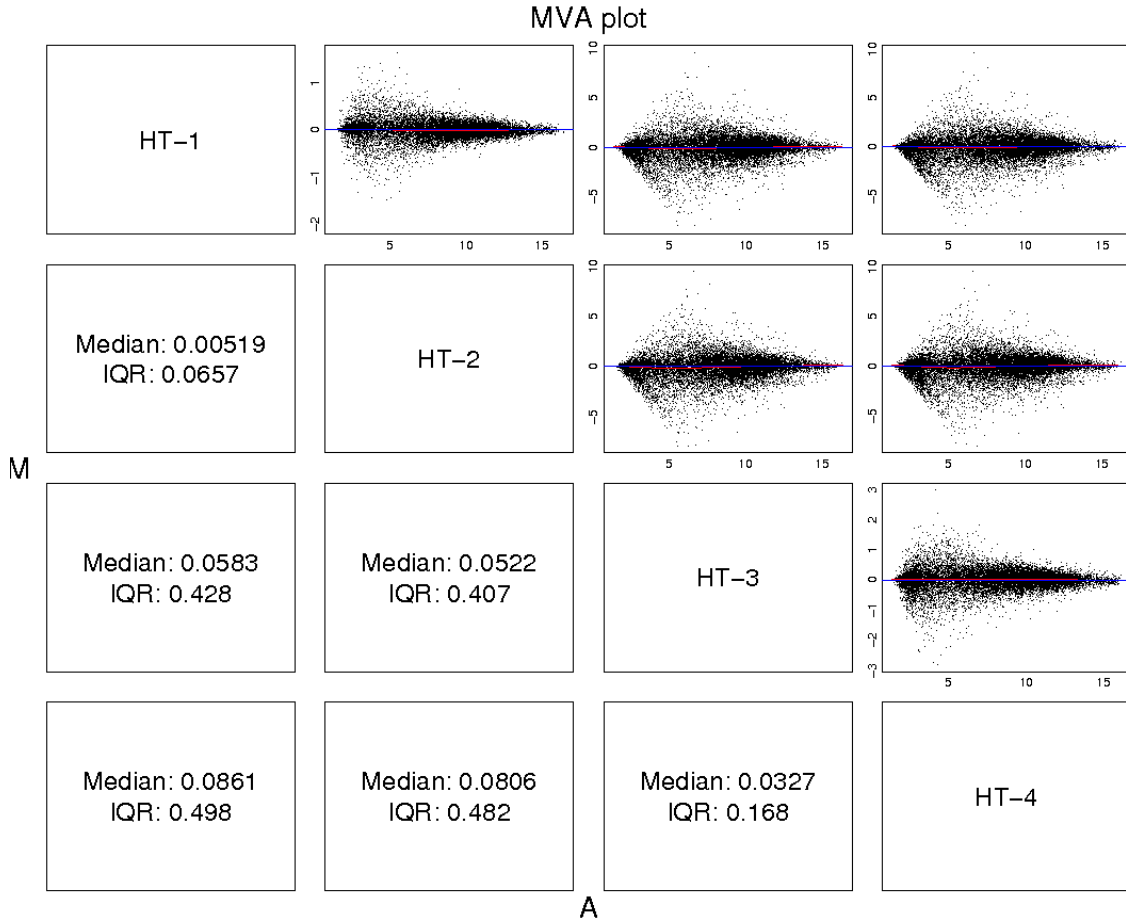
1. Array Quality Analyzer.
2. Differentially Expressed Gene Finder.
3. Cluster Pre-Processor.
4. Gene Ontology based group predictor and heatmap Creator.

Array Quality Analyzer

In any microarray data analysis, the arrays should be first analyzed for their quality. Array Quality Analyzer is written in R to find arrays of bad quality. Array Quality Analyzer does a pairwise comparison of the arrays and gives a plot of log ratio versus average log intensity of the arrays with the median and interquartile range (IQR) of each pair, this kind of plot is called an MVA plot. In an MVA plot of related arrays, the interquartile range should be small and the cloud points should be around zero. Arrays with an abnormal cloud of points and interquartile range are usually thrown out before the actual microarray data analysis.

In Figure 2 a MVA plot of four *C.elegans* controls given by the Array Quality Analyzer is shown.

Figure 1. MVA Plot of the Controls



Differentially Expressed Gene Finder

Genes that show difference in their expression value with reference to the control are called differentially expressed genes. The Differentially Expressed Gene Finder written in R finds the differentially expressed genes. Finding the differentially expressed genes is a multiple step process. Each of the steps involved are detailed below.

Normalization

In microarray experiments there are lot of experimental and systematic variations that can affect the accuracy of gene expression levels, e.g. differences in labeling, sample preparation, hybridization etc [3]. The idea of normalization is to adjust for the effects, due to variations and limitation of the technology involved.

Over the years a lot of algorithms have been introduced to normalize the microarray data. Some of the more popular ones are MAS [4], Robust Multichip Averaging (RMA) [5], Probe-level Linear Model fits (PLM) [6] and GCRMA an improvement in Robust Multichip Averaging algorithm (Using DNA sequence information to improve Robust Multichip Averaging). There is a wealth of literature in computational statistics, comparing each of the above method with each other [7] [8] [9] and it appears that the GCRMA algorithm performs better than the others in most of the cases studied. The reason for better performance of GCRMA is attributed to the fact that the algorithm takes into consideration the “GC” content in microarray data that is normalized.

Taking into account all the above studies, the Affymetrix microarray data is first normalized using the GCRMA algorithm by the Differentially Expressed Gene Finder.

Building a Linear Model

The second step in identifying the differentially expressed genes is building a linear model. Linear model building requires two matrices. The first matrix is called the

design matrix and each row in the design matrix corresponds to an array in the experiment and each column corresponds to a gene in the experiment. The second matrix is called the contrast matrix which specifies comparisons that need to be made between the samples. Once the above two matrices are made, the linear model is constructed using the `lmfit` method in the `limma` package [10] of R.

Calculating Differences in Gene Expression

Once, the linear model is built, the differences in expression levels of genes, across different arrays are calculated using the `contrasts.fit` method in the `limma` package. The arrays that are to be compared were specified earlier, in the contrasts matrix.

Identification of the Differentially Expressed genes

The number of ways to identify differentially expressed genes is overwhelming. In computational statistics, many studies have been done, in order to quantify the effectiveness of various methods [11]. As a result it has been accepted that the t-test is one of the better ways to identify differentially expressed genes. A series of t-tests are made using the `decideTests` method in the `limma` package, on the Affymetrix data, to identify differentially expressed genes with p value less than 0.05.

Cluster Pre-Processor

The Cluster Pre-Processor is a PERL based program that mainly has two functions. Its first function is to convert the Affymetrix gene names to the commonly accepted gene names and its second function is to center the gene expression data with respect to the controls for clustering. Once the pre-processing is done the actual clustering is done using Cluster [12] and the clustograms created are viewed and analyzed using jTreeView [13].

Gene Ontology Based Group Predictor and Heatmap Creator

Gene Ontology based group predictor and heatmap Creator are new Class Prediction methods based on the Gene Ontology database. Gene Ontology (GO) is a set of annotations of genes organized in three ontologies: biological process, cellular component and molecular function. The GO ontologies provide controlled vocabulary for annotating genes with annotation terms organized in the three hierarchical ontologies [14]. 14, 535 *C. elegans* genes have GO annotations. The Gene Ontology based group predictor and heatmap Creator programs annotate genes to particular GO terms or their child terms to create GO gene lists. The average expression change of the genes in each GO gene list is calculated and a p-value for the average expression change is calculated using Monte Carlo sampling. The heatmaps are then created using the heatmap creator for the GO groups. The heatmaps show the level of a expression of particular gene group.

Red indicates decreased expression for the gene group and blue indicates increased expression. Then the groups are hierarchically clustered using a modified version of heatmap.2 from the R statistics gplots package.

Kentucky MicroArray Analysis Suite vs other Microarray Analysis Software

The Kentucky MicroArray Analysis Suite fills a void in the area of Microarray data analysis as it offers an Array Quality Analysis tool, a tool to find differentially expressed genes, a precluster analysis tool and a Gene Ontology based analysis tool - all in one place. Since it is not proprietary there is no huge license fee like GeneSpring and it can be used freely almost anywhere. It has a novel Gene Ontology based analysis tool that helps Biologists and others to see the big picture. KMAS can easily be integrated into any existing microarray database, a feature that almost all microarray data analysis tools lack.

CHAPTER FOUR

RESULTS

Affymetrix Microarray data related to aging in *C. elegans* was analyzed using the Kentucky Microarray Analysis Suite (KMAS). The results of the analysis led to some interesting findings on genes related to aging in *C.elegans*.

Caenorhabditis elegans

C. elegans is a very common nematode found in terrestrial habitats feeding on micro-organisms. Sydney Brenner [15] started research into the biology of *C. elegans* around 1965. The organism is cheap and easy to maintain in a laboratory, hence widely used as a model organism. The *C. elegans* genome project was completed in 1998 [16] making *C. elegans* the first multicellular organism whose genome was fully sequenced.

Reproduction in *C. elegans*

C. elegans exists as males and hermaphrodites in which both male and female sex organs are present. The hermaphrodites can produce both sperm and egg; hence it can reproduce by self fertilization. The male produces only sperm which compete with sperm produced by the hermaphrodite after mating to fertilize the egg of a hermaphrodite. It has been observed that cross-fertilization produces males and hermaphrodites in equal proportions but self-fertilization produces only hermaphrodites. Males and hermaphrodites both have the same general anatomy.

Life cycle of *Caenorhabditis elegans*

The eggs are laid by the hermaphrodite. Once the eggs hatch, they pass through four larval stages called L1 to L4. In a crowded or in the absence of food condition, *C. elegans* can enter an alternative larval stage called the dauer state. The dauer larva are stress-resistant and do not age. The hermaphrodites produce sperm during the L4 stage, and lay eggs as adults. At 20 °C the laboratory strain of *C. elegans* has an average life span of approximately 2-3 weeks and a generation time of approximately 4 days.

Evolutionary Relationship

On the basis of evolution in the animal kingdom the Nematodes are considered to have a deep branch, close to the separation of insect and vertebrate lineages. Many of the basic mechanisms needed for functioning of a complex organism may have evolved before the time nematodes branched away from rest of the animal kingdom. Genes have been fairly well conserved to such an extent that in many experiments *C. elegans* genes have been substituted for vertebrate genes. Around 84 human disease genes have been cloned and in that set 25 have a direct ortholog in *C. elegans* and 43 have a very good similarity with other *C. elegans* orthologs [19]. *C. elegans* due to its similarity and evolutionary relationship with other nematodes is also a very good model organism in the studies of parasitic nematodes which are of major significance in medicine and agriculture.

Genome of *Caenorhabditis elegans*

The *C. elegans* genome was sequenced by a consortium based at the Sanger Center, Cambridge, UK and the Genome Center based at the Washington University in St. Louis, USA. Since it was the first multicellular organism to be sequenced it served as a model for the Human Genome Project. The *C. elegans* genome is 97 Mega bases in size, about 30 times smaller than the human genome. It is estimated that *C. elegans* have about 20,000 protein-coding genes. The *C. elegans* genome is spread across six approximately equally sized chromosomes, five autosomes (non-sex chromosomes) and one X chromosome.

Mitochondria

The mitochondria are called the powerhouse of the cell. They have two membranes. The first membrane limits and encloses the mitochondrion and the second membrane folds inwards. The Mitochondria convert oxygen and nutrients to adenosine triphosphate (ATP) which powers the cell's metabolic activities, this process is called aerobic respiration.

Mitochondria and Aging

Mitochondria have their own DNA and the mitochondrial DNA is quite different from nuclear DNA. It exists as a simple plasmid and it is similar to bacterial DNA rather than the chromosomal DNA of higher organisms. It is not associated with histones, and

the DNA repair mechanisms that correct damage in nuclear DNA are missing from mitochondrial DNA. The electron transport system is not perfect in mitochondria as a result electrons interact with the oxygen and produce superoxide radicals (or hydroxyl radicals). Due to the proximity of the mitochondrial DNA to the superoxide radicals and its lack of protection and repair mechanisms, leads to free radical-mediated mutations and deletions in the mitochondrial DNA occur which causes mitochondrial ageing and degeneration.

Mitochondrial Genes

The *C. elegans* mitochondria has 284 genes, out of these, 227 genes have been annotated by WormMart, and an additional 57 have been annotated by the Gene Ontology database. According to the Kyoto Encyclopedia of Genes and Genomes [17], of the 284 mitochondrial genes 93 are known to be part of several metabolic pathways.

Aging

The process of aging has many definitions. Aging can be defined as a biological process that increases the chance of death as an animal gets chronologically older. Aging can also be defined in terms of changes that an animal undergoes as it ages. As an animal gets older, not all changes may be relevant to aging. Some may not result in an increased chance of death. Some changes may be causes of aging while others may be only consequences, or markers of aging.

Effects of Aging in Humans and *C. elegans*

Aging affects most tissues of the body. In humans, the features of aging include graying and loss of hair, reduced subcutaneous fat, cardiovascular disease, decalcification of the bones, loss of neurons etc., in general as well as specific populations. Every organ of the body shows reduced functioning, including the immune system, the kidney, the heart, and the liver. The cellular phenotype includes shortening of the telomeres and auto-fluorescent deposits. Interestingly, in old individuals, stem cells make up a larger proportion of the bone marrow, and are dividing more actively than in the young. Cells from older individuals divide fewer times in culture before senescing. In *C. elegans*, features associated with aging include torpor and finally cessation of movement, cessation of reproduction, accumulation of auto-fluorescent deposits in cells. All somatic cells in the adult worm are post-mitotic, and therefore telomere shortening likely does not play a role in aging. In physiological and cellular level aging has very similar effects on both humans and *C. elegans* such as declining organ function, weakened stress resistance, decrease of mitochondrial function, lipofuscin deposits and accumulation of misfolded or damaged proteins.

Aging and Mitochondrial Genes

RNAi knockdown of complex III gene *cyc-1* [20], complex I gene D2030.4 [24], *aco-2* [21], *phb-2* [22], T22B11.5 and reduced expression of *frh-1* [23] have been shown to increase the lifespan of *C. elegans*.

The effect of RNAi gene expression knock-down of these mitochondrial genes on genomic gene expression is examined. The genes expression changes in the RNAi samples were analyzed to determine the systemic effects on gene and their relation to ageing in *C.elegans*.

Microarray Experimental Design

RNAi knockdown of the selected mitochondrial genes was induced by feeding the worms bacteria containing dsRNA vectors. The bacteria was plated on NGM plates with IPTG to induce dsRNA expression for 16 hours. Synchronized *rrf-3* strain L1 worms were added to the plates (3x10⁵ worms/15cm plate). Worms were transferred to fresh RNAi plates containing FUDR the first day of adulthood, day 3. On the third day of adulthood, the strains were collected and RNA was extracted at 4°C using a standard Trizol protocol. RNA was purified by spin column (Qaigen) and hybridized to Affymetrix arrays at the University of Kentucky Microarray Core facility. The Affymetrix *C.elegans* Genechips contain 22,625 probe sets and cover 90% of *C. elegans* genes. Two biological replicates were done for the control and four were done for each RNAi knockdown. These experiments were carried out in Dr. Jim Lund's lab at the University of Kentucky by Mr. Scott Frasure.

Analysis of Differentially Expressed Genes

The Differentially Expressed Gene Finder program in the KMAS suite, identified 465 unique genes that were differentially expressed (upregulated or down regulated) in at least one mutant and there was a substantial overlap between the significantly expressed genes from all the six mutants. The *frh-1* and *phb-2* mutants had the fewest differentially expressed genes among the six mutants.

Table 1. Microarray expression analysis of mitochondrial genes
KMAS found 465 unique genes were either up or down regulated.

Gene Knockouts	Number of genes upregulated	Number of genes downregulated
<i>cyc-1</i>	164	186
D2030.4	270	406
<i>Aco-2</i>	206	176
<i>frh-1</i>	56	32
<i>Phb-2</i>	58	30
T22B11.5	270	324

Classification of the Differentially Expressed Genes

WormMart – the annotation database for *C.elegans*, had annotations for 366 of the 465 unique differentially expressed genes and these 366 genes were classified into categories based on their function. Out of the 366 genes, 144 had an unknown function, 98 genes were involved in protein synthesis, and 60 genes were involved in cell growth and defense.

Table 2. Classification of differentially expressed genes based on function
 Out of the 465 unique differentially expressed genes, 366 genes were classified into categories based on their function.

Functional Class	Number of Genes
Kinases	9
Cellular Transporters	17
Signaling	4
Protein Synthesis	98
Metabolism	34
Cell Growth and defence	60
Unknown	144

Clustering of Differentially Expressed Genes

The 465 differentially expressed genes were pooled together using the Cluster Pre-Processor, and hierarchically clustered by Euclidean distance using single linkage. Hierarchical clustering by experiment (RNAi gene) shows two main clusters, one of D2030.4, *cyc-1*, *aco-1* and T22B11.5 and the other of *frh-1* and *phb-2*. The *frh-1* and *phb-2* clusters show similar gene expression changes but with reduced magnitude. These two samples formed a subcluster in most gene classes examined and often showed a greater differentiation from the other four samples.

Gene Ontology Analysis of the Gene Expression Data

Gene Ontology based group predictor and heatmap Creator program of KMAS, is used to classify genes into groups based on their annotation. Once the genes are classified into groups the heatmap creator creates heatmaps for each GO group.

In the Molecular function heat map, the D2030.4, *cyc-1*, and *aco-1* samples show increased expression of genes that hydrolyse O-glycosyl compounds indicating a metabolic shift to increased glycogen breakdown. Also, the structural molecular activity GO group shows decreased expression in *phb-2* and *frh-1* mutants. In the Cellular Component heat map, the T22B11.5 mutant showed increased expression levels in Integral to membrane, Intrinsic to membrane, and Membrane GO groups which may indicate large changes in the composition of cellular membranes. In the Biological process heat map the embryonic development and related GO groups shows decreased expression in four of the six mutants indicating a major decrease in reproductive capacity. The *frh-1* and *phb-2* samples show a significant increase in expression of genes involved in carbohydrate catabolism indicating a decline in inputs to the respiratory metabolism.

Figure 2. Clustergram of pooled differentially expressed genes from the six mitochondrial gene RNAi.

Hierarchical clustering by experiment shows two main clusters, one of D2030.4, *cyc-1*, *aco-1* and T22B11.5 and the other of *frh-1* and *phb-2*.

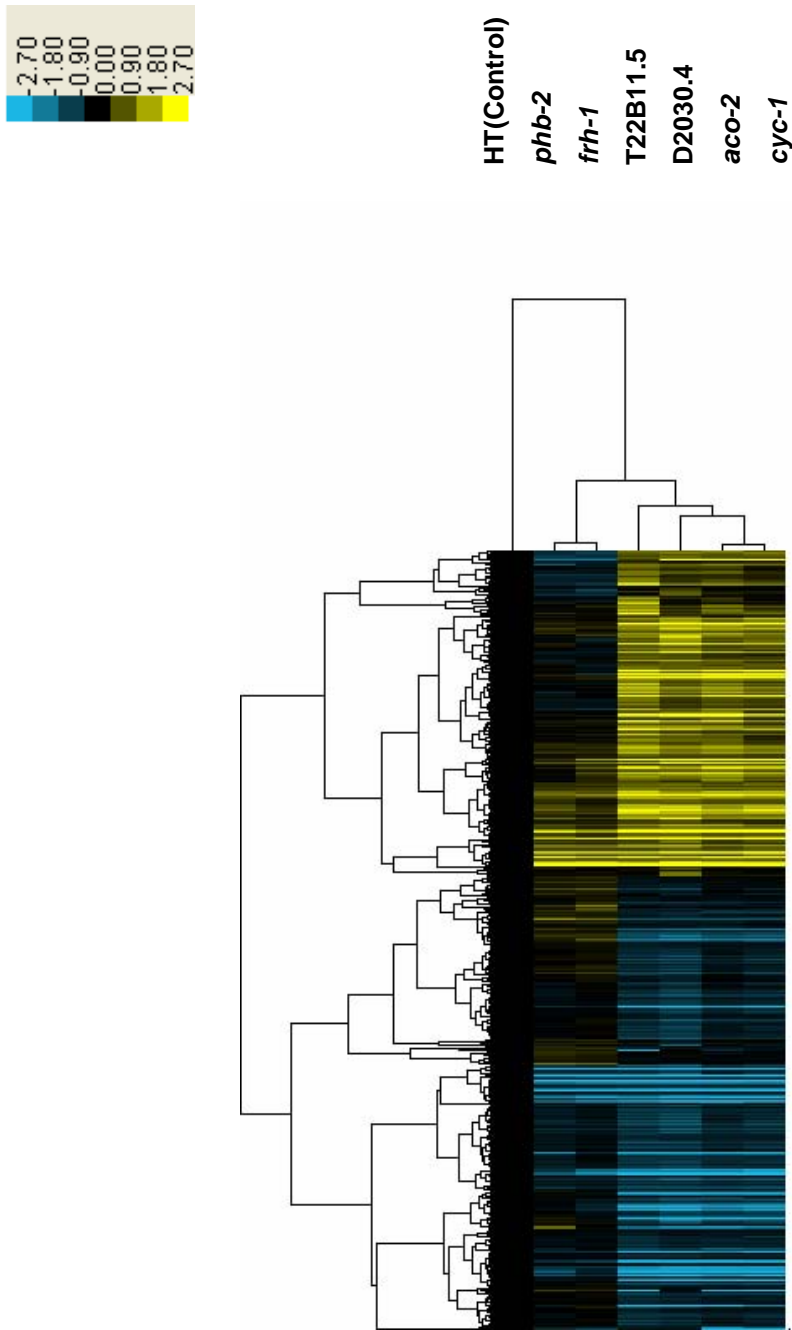


Figure 3. GO Molecular Function heat map

In the Molecular function heat map, the D2030.4, *cyc-1*, and *aco-1* samples show increased expression of genes that hydrolyse O-glycosyl compounds. Also, the structural molecular activity GO group shows decreased expression in *phb-2* and *frh-1* samples.

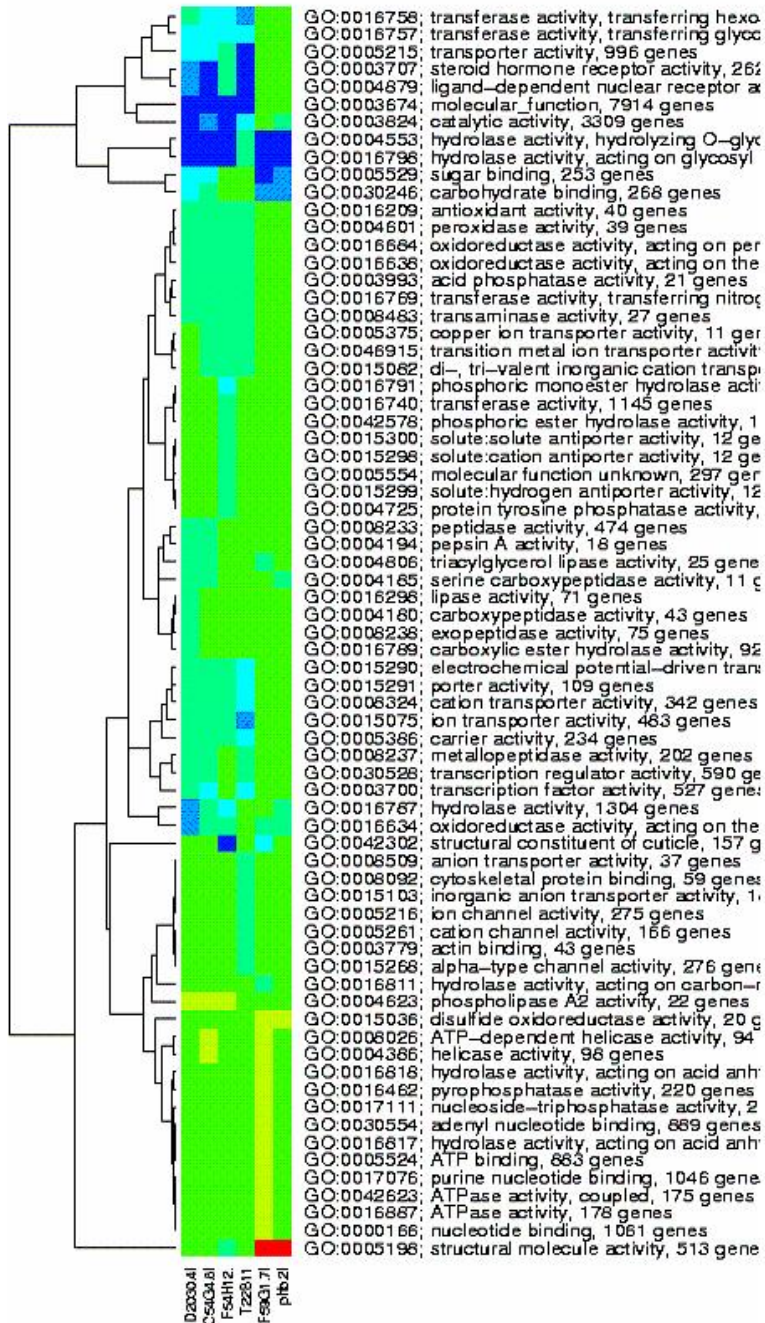


Figure 4. GO Cellular Component heat map
 The T22B11.5 sample showed increased expression levels in Integral to membrane, Intrinsic to membrane, and Membrane GO groups.

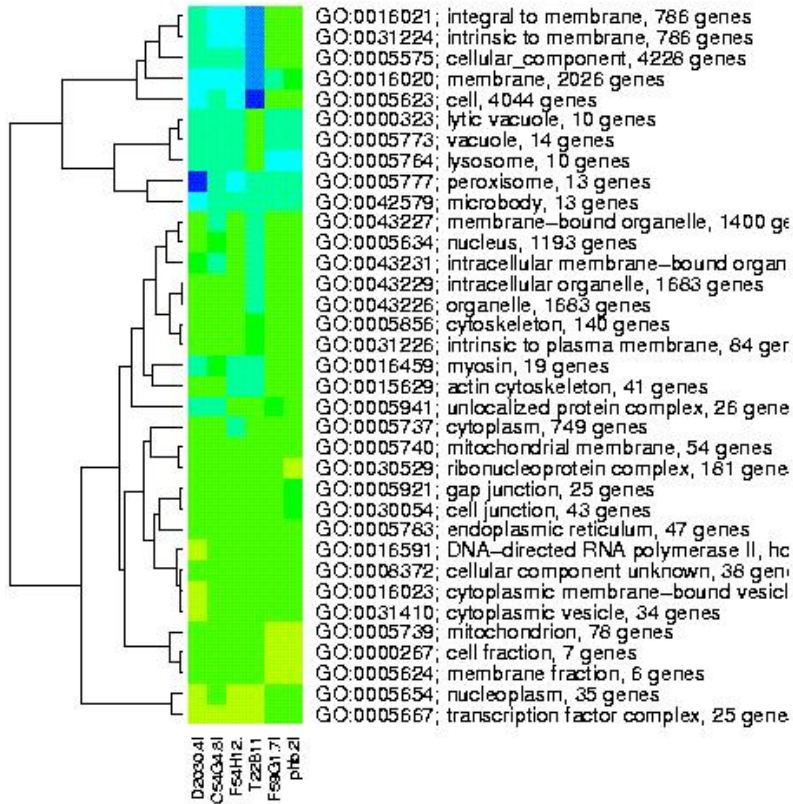
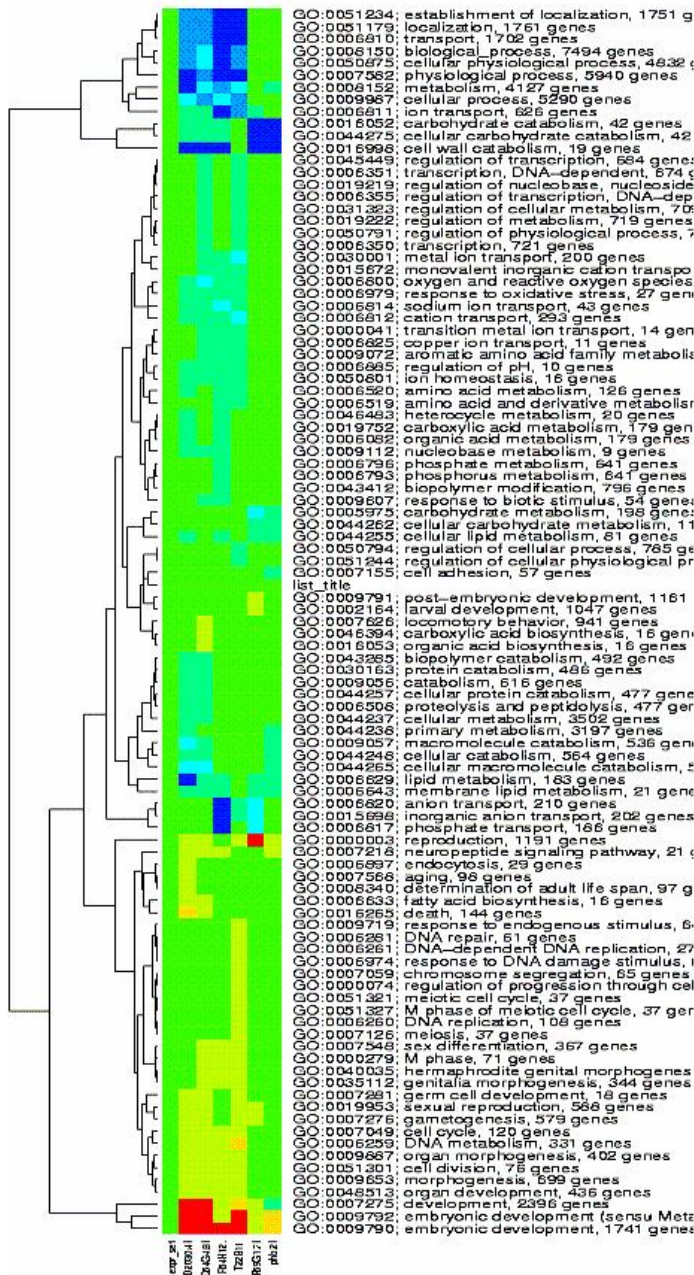


Figure 5. GO Biological Process heat map

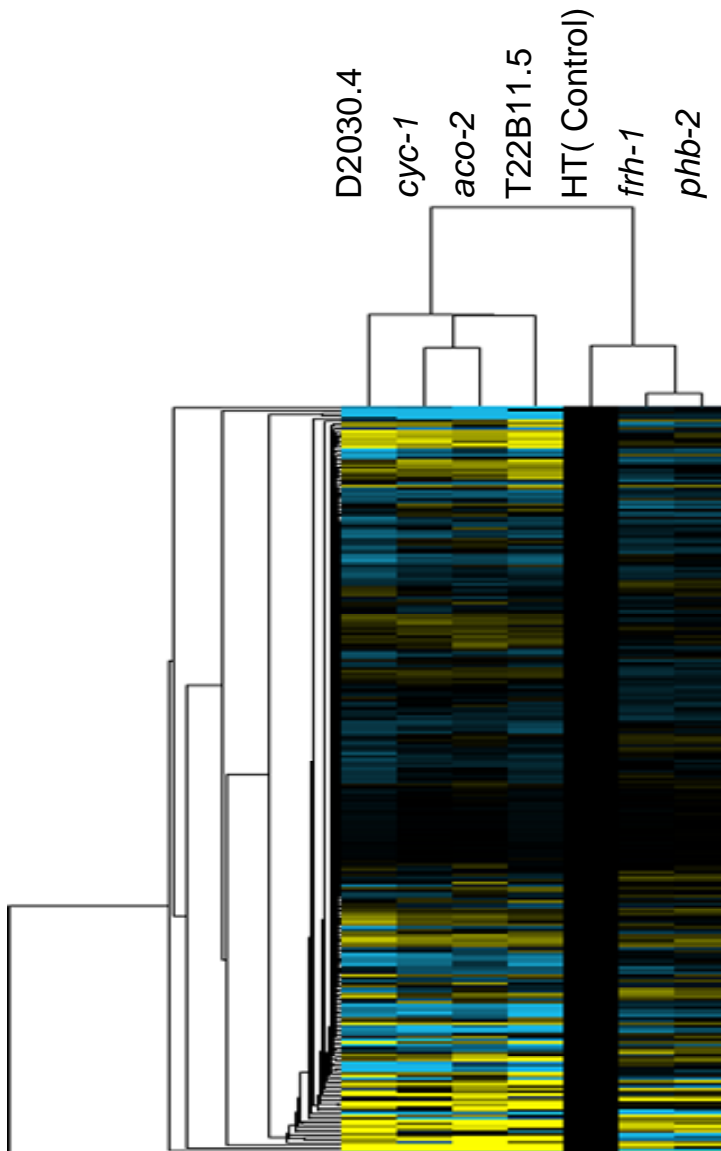
In the Biological process heat map the embryonic development and related GO groups shows decreased expressed in four of the six mutants. The *frh-1* and *phb-2* samples show a significant increase in expression of genes involved in carbohydrate catabolism.



Mitochondrial Gene Expression

A list of 284 mitochondrial genes was assembled. Out of these 284 genes, 227 were from WormMart and an additional 57 were from the GO. The expression values for these 284 genes were extracted from the linear statistical model constructed earlier using Differentially Expressed Gene Finder. These genes were pooled together using the Cluster Pre-Processor. When a hierarchical clustering of the 227 mitochondrial genes in the six mutants was done, substantial similarities in expression patterns were observed in four of the six mutants. Two of the mutants, *frh-1* and *phb-2*, showed similar levels relative to the control HT115 sample are yellow and genes with reduced expression are blue expression patterns which formed a subcluster.

Figure 6. Clustogram of 284 mitochondrial genes
In a hierarchical clustering of 227 mitochondrial genes in all six samples, substantial similarities in expression patterns were observed in four of the six samples. Two of the samples, *frh-1* and *phb-2*, showed similar expression patterns which formed a subcluster.



CHAPTER FIVE

CONCLUSIONS

The Kentucky Microarray Analysis Suite helped in the analysis of Affymetrix microarray data related to aging in *C.elegans*, using three different approaches. This kind of novel method of analysis leads to many interesting conclusions.

Of the 465 differentially expressed up or down regulated genes, 98 genes are involved in protein synthesis, and 34 genes were involved in metabolism. Hence, RNAi knockdown of mitochondrial genes can have a significant impact on protein synthesis.

A Gene Ontology based analysis of 14,535 genes using the Gene Ontology based group predictor and heatmap Creator showed an increased expression of genes that hydrolyze O-glycosyl compounds indicating a metabolic shift to increased glycogen breakdown. Increased expression levels were also observed in the in the Integral to membrane, Intrinsic to membrane, and Membrane GO groups which may indicate large changes in the composition of cellular membranes. Embryonic development; and related GO groups shows decreased expression in four of the six mutants indicating a major decrease in reproductive capacity and *frh-1* and *phb-2* mutants show a significant increase in expression of genes involved in carbohydrate catabolism indicating a decline in inputs to the respiratory metabolism.

Overall, *frh-1* and *phb-2* samples show a significantly different expression pattern than the rest of the mutants. The number of significantly expressed genes in these two samples is considerably less, when compared to the other four samples.

The study of gene expression patterns in the six mitochondrial mutants has shown that RNAi knockout of the selected mitochondrial genes has a significant impact on the gene expression patterns of numerous other genes. The change in expression patterns cause changes in metabolism, protein synthesis, and reproduction. These changes seem to increase the lifespan of *C. elegans*.

Future Work

The expression levels of the list of genes identified as differentially expressed need to be measured by some biological method like Real Time PCR. The methods in Kentucky Microarray Analysis Suite can be extended to other popular Microarray formats like Agilent, and integrated into the standard Stanford Microarray Database template so that it is useful to a large audience.

REFERENCES CITED

1. Dobbin K, Simon R - Comparison of microarray designs for class comparison and class discovery. (Bioinformatics. 2002 Nov;18(11):1438-45.)
2. Somorjai RL, Dolenko B, Baumgartner R. - Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. (Bioinformatics. 2003 Aug 12;19(12):1484-91.)
3. Normalization of cDNA Microarray Data (METHODS: Selecting Candidate genes from DNA Array Screens: Application in Neuroscience)
4. The MAS 5.0 User Manual
(http://www.affymetrix.com/Auth/support/downloads/manuals/mas_manual.zip)
5. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). (Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264. 4/2003)
6. Bolstad B. affyPLM: Methods for fitting probe level models to Affy data.
<http://www.bioconductor.org/repository/devel/vignette/affyPLM.pdf>
7. Bolstad B. Why do my MAS 5.0 values differ? <http://stat-www.berkeley.edu/~bolstad/MAS5diff/Mas5difference.html>
8. Arnold Saxton - A Comparison of Expression Measures from Affy GeneChip Arrays. (KBRIN, Lake Barkley 2006)
9. Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. (Bioinformatics, *19*, 185, 2003).
10. Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. (Statistical Applications in Genetics and Molecular Biology, Vol. 3, No. 1, Article 3)

11. Gordon K. Smyth, Yee Hwa Yang and Terry Speed - Statistical Issues in cDNA Microarray Data Analysis (Functional Genomics: Methods and Protocols, M J. Brownstein and A. B.Khodursky (eds.), Methods in Molecular Biology series, Humana Press, Totowa, NJ, 2002)
12. Eisen et al. (1998) PNAS 95:14863.
13. Saldanha, 2004.
14. <http://geneontology.org>.
15. <http://elegans.swmed.edu/Sydney.html>.
16. C. elegans Sequencing Consortium - Genome sequence of the nematode C. elegans: a platform for investigating biology, Science. 1998 Dec 11;282(5396):2012-8.
17. <http://www.genome.jp/kegg/>.
18. Affymetrix GeneChip® Manual.
19. Mushegian et al Genome Research 1998.
20. Dillin et al 2002.
21. Hamilton et al 2005.
22. Artal-Sanz et al 2005.
23. Ventura et al 2005.
24. Lee et al 2003.