2007

A VISUALIZATION TOOL FOR CROSS-EXPERIMENT GENE EXPRESSION ANALYSIS OF C. ELEGANS

Lin Xue
University of Kentucky, lionetxue@gmail.com

ABSTRACT OF THESIS

A VISUALIZATION TOOL FOR CROSS-EXPERIMENT GENE EXPRESSION ANALYSIS OF *C. ELEGANS*

Forty-six genomic gene expression studies of free living soil nematode *C. elegans* have been published. To facilitate exploratory analysis of those studies, we constructed a database containing all the published *C. elegans* expression datasets. A Perl CGI program, called Microarray Analysis Display (MAdisplay), allows gene expression clustergrams of any combination of entered genes and datasets to be viewed (http://elegans.uky.edu/gl/madisplay). Perl programs were used to preprocess the raw data from different sources into a common format and to transform the data to display the expression changes relative to each experiment's controls. Three hundred lists of genes from figures and tables were extracted from the publications and made available in the GeneLists database, which also contains Gene Ontology and KEGG gene lists. We used these tools to examine in a systematic fashion the mean expression of gene lists in the set of microarray and SAGE experiments. Seventy-nine percent of publication derived gene lists show a strong expression change (p-value <0.001) in more than one experiment with the median being fourteen out of the 127 experiments that are derived from the forty-six publications. This indicates that groups of genes identified in one publication typically show an expression effect in many other experiments.

KEYWORDS: *C. elegans*, gene expression, microarrays, Gene Ontology, clustering

Lin Xue

7/11/2007

A VISUALIZATION TOOL
FOR CROSS-EXPERIMENT GENE EXPRESSION
ANALYSIS OF *C. ELEGANS*

By

Lin Xue

Jim Lund
Director of Thesis

Brian Rymond
Director of Graduate Studies

7/11/2007
Date

A VISUALIZATION TOOL
FOR CROSS-EXPERIMENT GENE EXPRESSION
ANALYSIS OF *C. ELEGANS*

By

Lin Xue

Chuck Staben
Co-Director of Thesis

Jerzy Jaromczyk
Co-Director of Thesis

Brian Rymond
Director of Graduate Studies

7/11/2007
Date

RULES FOR THE USE OF THESES

Unpublished theses submitted for the Master's degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgments.

Extensive copying or publication of the thesis in whole or in part also requires the consent of the Dean of the Graduate School of the University of Kentucky.

A library that borrows this thesis for use by its patrons is expected to secure the signature of each user.

Name Date

_____

_____

_____

_____

_____

_____

_____

_____

THESIS

Lin Xue

The Graduate School

University of Kentucky

2007

A VISUALIZATION TOOL FOR CROSS-EXPERIMENT GENE EXPRESSION ANALYSIS OF *C. ELEGANS*

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in the
College of Arts and Sciences
at the University of Kentucky

By

Lin Xue

Lexington, Kentucky

Director: Dr. Jim Lund, Assistant Professor, Department of Biology

Lexington, Kentucky

2007

A VISUALIZATION TOOL FOR CROSS-EXPERIMENT GENE EXPRESSION ANALYSIS OF *C. ELEGANS*

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in the
College of Arts and Sciences
at the University of Kentucky

By

Lin Xue

Lexington, Kentucky

Co-Directors: Dr. Chuck Staben, Associate Professor of Department of Biology
and Dr. Jerzy Jaromczyk, Associate Professor of Department of Computer Science

Lexington, Kentucky

2007

MASTER'S THESIS RELEASE

I do not authorize the University of Kentucky
Libraries to reproduce this thesis in
whole or in part for purposes of research.

Signed: Lin Xue

Date: 7/11/2007

ACKNOWLEDGEMENTS

The following thesis, while an individual work, benefited from the insights and direction of several people. First, my advisor and thesis chair, Dr. Jim Lund, has helped me to gain knowledge and experience in bioinformatics. His constant friendship and instructive comments accompanied me at every stage of the thesis process as well as my whole graduate study. I appreciate you, Jim, for exemplifying the nice, smart, open-minded intellect to which I aspire. I should also like to thank the other members of my dissertation committee, Dr. Chuck Staben and Dr. Jerzy Jaromczyk. Both provided insights that guided and challenged my thinking, polishing the finished project. I should especially thank Dr. Staben for including me in his research as a rotation student in the first semester. You have all helped make my experiences here at the University of Kentucky very fulfilling.

Appreciation is expressed to fellow colleagues in Dr.Lund's Lab: Scott Frasure (for being my English teacher, big brother and a real delight to everyone in the lab), Vijay Raghavan (for all your help with computer problems), Tseten Yeshi (for intelligent comments in lab meetings and same bright jokes off meetings), George Chaffins (for yummy snacks), Suchita Desai (for girl talks), and Brandon Barker (for sharing the same interest of bioinformatics with me and supporting my dreams). Thank you all for providing support throughout the project, as well as enriching my pleasurable personal

time in graduate school.

Additionally, I received equally important assistance from family and friends. I especially thank you, Mom and Dad, for bringing me up and support all my career decisions, including flying to the opposite of the earth to pursue higher education. I should like to acknowledge my cousin, Sheng, who has been my best pal since childhood and has been the cornerstone of my support system since we both came to the US.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1 Introduction

A technological revolution in the 1990s resulted in the completion of genomic sequences of several model organisms, including *Saccharomyces cerevisiae* (yeast) [Goffeau A et al. 1996], *Drosophila melanogaster* (fruit fly) [Adams MD et al. 2000], and *Caenorhabditis elegans* (nematode worm) [The *C. elegans* Sequencing Consortium, 1998]. This has lead biologists to seek to translate genomic sequence information into functional biological mechanisms that will allow researchers to gain an understanding of the control of genes during normal growth and development as well as in disease states.

The development of high throughput transcription profiling technologies such as cDNA microarrays [Drmanac S, 1996], oligonucleotide arrays [Lockhart DJ, 1996] and Serial Analysis of Gene Expression (SAGE) [Velculescu VE et al. 1995] have been designed to tackle the task of genomic scale analysis of gene expression. Microarrays have become an essential tool in gene expression research. Spotted microarrays (or two-color or two-channel microarrays) use oligonucleotides, cDNA or PCR products that correspond to mRNA as probes. These probes are spotted onto the microarray chip surface, typically glass. Typically cDNA from the two samples to be compared is labeled with two different fluorophores then mixed and hybridized to a single microarray. Gene expression differences between the two samples can then be read from their relative intensities using a special purpose confocal scanner.

Affymetrix GeneChips are composed of multiple probes designed to match parts of the sequence of known or predicted mRNAs. The *C. elegans* GeneChip is a whole-genome array designed to assay over 22,500 transcripts from almost 19,000 genes. (http://www.affymetrix.com/products/arrays/specific/celegans.affx) Each transcript is measured by 11 probe pairs, which consist of a perfect match 25mer oligonucleotide (PM) and a 25mer mismatch oligonucleotide (MM) that contains a single base pair mismatch in the central position. The PM/MM design is used for identification and subtraction of nonspecific hybridization and background signals. The oligonucleotides are synthesized on the array surface using a photomasking process allowing high density arrays to be constructed with very low variation in oligo density. Affymetrix GeneChip arrays have standardized array fabrication and processing protocols that result in low technical variability and good reproducibility.

Serial analysis of gene expression (SAGE) is another way to perform global profiling of gene transcripts. It involves the generation, concatenation and sequencing of short 14–21 base pair cDNA segments (tags) that typically uniquely correspond to expressed sequences [Pleasance ED et al 2003]. Unlike microarray technologies, SAGE does not require a priori knowledge of the genes to be analyzed. Moreover, SAGE gives a potentially unbiased sampling while microarray data depends on the experimental state of gene models. A drawback of SAGE is that the cost of is higher than for microarrays. The raw SAGE data is counts of cDNA tags and the relative

proportions of tags from different genes reflect the abundance of their mRNA level in the sample. Thus, statistical significance relies on the depth of tag coverage.

Microarrays and SAGE have enabled the exploratory analysis of gene expression on a genome wide scale. In particular, it has made possible experiments in the model organism *Caenorhabditis elegans*, a free living soil nematode with well-developed genomic resources, to monitor expression levels of the full set of genes simultaneously and quantitatively.

C. elegans Gene Expression Studies

Gene expression data have been produced from hundreds of experiments using *C. elegans* and has examined various stages, specific tissues, different sexes and mutants, and various environmental conditions. There are several reasons to choose the tiny nematode for experimental studies.

First, it is the simplest multicellular organism whose genomic sequence being completed [*C. elegans* Sequencing Consortium 1998], and it is the only multicellular organism whose cell lineage is completely documented [Sulston and Horvitz 1977; Sulston et al. 1983].

C. elegans has a short life span of 2 weeks and generation time of just 3 days which allows much shorter and more cost-effective experimental procedures than studies on larger animals. Its small size, transparency, and limited number of cells make it a good subject to observe many complex cellular and developmental processes that cannot easily be observed in more complex organisms.

One central advantage of performing microarray analysis in *C .elegans* is the ease of RNAi experiments makes it amenable to both forward and reverse genetics. Knock-down of the target genes by RNA interference will produce strong mutant phenocopies and can be used either as a way to produce mutant strains for microarray analysis or as a confirmation test to validate the genes identified in microarray experiments.

Stuart Kim is the pioneer in the development of DNA microarrays for *C. elegans* gene expression profiling [Astin J at al. 2004]. The first DNA microarray paper published by Kim's lab identified 1416 germline-enriched transcripts that define three groups: the sperm-enriched group, the oocyte-enriched group and the germline-intrinsic group, defined as genes expressed similarly in germlines making only sperm or only oocytes [Reinke V et al. 2000]. Later, they devised a sample preparation method to isolate mRNA in cells or tissues of interest, called messenger RNA tagging. The basis of the technique is to use an epitope-tagged poly (A)-binding protein (PAB-1), expressed under the control of a muscle-specific promoter, to co-precipitate mRNAs preferentially enriched in muscle. [Roy et al. 2002] Another commonly used tissue specific sampling method is to mark target cells or tissues with green fluorescent protein (GFP), followed using fluorescence-activated cell sorting. This method is widely applied (seven

out of eight neuronal publications in our database) in neuronal research to either isolate neuron-specific mRNA or confirm the expression location of target genes.

A. A. Hill and his coworkers [Hill et al. 2000] were the first group to use a commercially-generated oligonucleotide based array in an experiment examining changes in transcriptional patterns during development and aging. The notable strongpoint of such arrays over spotted arrays is that they quantify mRNA levels while spotted arrays only measure relative transcript levels.

As a complementary method, SAGE analysis is able to identify expression changes not detected in related experiments using DNA microarrays. It is more expensive to perform as it requires high-capacity DNA sequencing, and thus is the least used technique of gene expression analysis (35 out of 881 arrays in our database). The Genome Sequencing Center in British Columbia (BCGSC) at http://elegans.bcgsc.bc.ca is the source of all publicly available SAGE data. Their early work targeted differences in gene expression patterns of wild-type and dauer, a long-lived developmentally arrested larval stage in worms, and identified over 2000 dauer enriched genes. [Jones et al. 2001] Another significant publication by the same group used SAGE analysis on all developmental stages of intact animals and on selected purified cells and tissues of *C. elegans.* [McKay et al. 2003]

Technical Problems with Utilizing Published Data

With quite a bit of *C. elegans* gene expression data has being published access to that data is not convenient. A clear problem comes from different types of primary data collected using the different technologies. Spotted microarrays, oligonucleotide arrays, and SAGE each has its own features; different data analysis software is used to processes and normalize the data, and this produces raw data files in different formats. What is more, commercial Affymetrix chips have their own system of gene identifiers for the genes. Data was scattered in different resources, stored in variant formats, with diverse names or symbols for genes, thus reducing the power of searching. This situation motivated us to build an online freely accessible database for all the *C. elegans* gene expression data so that the biologists can quickly access and easily utilize the data regardless of its original format and derivation.

Gene Ontology and KEGG

The Gene Ontology project, or GO, provides a common controlled vocabulary to describe genes and gene products in any organism. Ontologies are specifications of a relational vocabulary. The use of ontologies facilitates making standard annotations and improves computational queries. While BLAST [Altschul SF, 1990] helps to search for homologs from different species based on similar sequences, GO helps to search for equivalent gene products from different species even though they may have

significantly different sequences or structures.

The Gene Ontology project consists of two main parts. The first is the ontology itself, made up of three categories, each representing a key concept in Molecular Biology: the molecular function of gene products; their role in multi-step biological processes; and their localization to cellular components. The second part is annotation; gene products are characterized using terms from the ontology. The members of the GO Consortium submit their data and it is made publicly available through the GO website at http://www.geneontology.org/, which is regularly updated with new versions of GO annotation files available for download on a monthly basis. [The Gene Ontology Consortium, 2000]

The controlled vocabularies of GO terms are hierarchical so that researchers can query them at different levels: for example, you can use GO to find all the gene products in the *C. elegans* genome that are involved in signal transduction, or you can zoom in on 'negative regulation of Ras protein signal transduction'. On the other hand, annotators can also take advantage of this hierarchical structure to assign properties to gene products at different levels depending on how detailed the knowledge is.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) at http://www.genome.jp/kegg/ is a knowledge base for systematic analysis of gene functions in terms of the networks of genes and molecules. It aims at computerizing the current knowledge of genetics, biochemistry, and molecular and cellular biology in terms of pathways of interacting molecules or genes. The major component of KEGG is the PATHWAY database that consists of graphical diagrams of biochemical pathways including most of the known metabolic pathways and some of the known regulatory pathways. The KEGG pathway database contains the information of how molecules or genes are organized in signaling networks and is complementary to most existing molecular biology databases that contain the information of individual molecules or individual genes. [Ogata H, 1999]

Clustering Analysis and Heat maps

The rapid advance of microarray and SAGE gene expression analysis has produced a huge amount of data and this invites development of new strategies to study this mass of biological data. Cluster analysis was developed to address the analysis of genome-scale experiments, allowing biologists to build a comprehensive understanding of the gene expression profiles being studied. Clustering the resultant expression data reveals groups of genes that share similar expression patterns, and based on the knowledge of a few known genes within those clusters, the role of novel genes in the same group can be speculated. For instance, Baugh and his colleagues found a set of 106 clusters representing a variety of very complex expression patterns in the *C. elegans*

early embryonic transcriptome and which reveal the most predominant expression patterns as well as significant associations of gene annotations. [Baugh LR 2003]

Clustering algorithms are based on the notion of unsupervised learning in which data objects within the same cluster are similar to one another and dissimilar to the objects in other clusters [Han and Kamber, 2001]. It is useful to think of the gene expression values in a microarray data set as a matrix, with each row being data for a single gene and each column being data for a single array/experiment. Each gene in the matrix defines a gene expression vector, which has as many dimensions as there are data points within the vector. Using standard mathematical metrics, the similarity (or dissimilarity) between different vectors can be then measured, and in conjunction with certain rules (an algorithm), these metrics can then be used to organize data. [Gollub J and Sherlock G, 2006]

Many different clustering methods have been developed. Those in most common use are hierarchical clustering, K-means clustering, and self-organizing maps (SOM). In the gene expression data exploration tools we describe here we use hierarchy clustering.

Hierarchical clustering builds or breaks up a hierarchy of clusters. The traditional representation of this hierarchy is a tree (called a dendrogram), with individual elements at the leaves and a single cluster containing every element at the root of the tree. The result of clustering is ordered data but the dataset is still as massive as the original observations. To make large data arrays easy for people to comprehend clusters are typically represented as a heat map, a graphical representation of data where values taken by a variable in a two-dimensional map are quantitatively and qualitatively reflected as colors. The end product of data clustering is a representation of complex gene expression data that, through statistical organization and graphical display allows biologists to assimilate and explore the data in a natural intuitive manner. [Eisen MB et al. 1998]

Databases and On-line Tools

WormBase (http://www.wormbase.org) is a comprehensive repository for information on *Caenorhabditis elegans* and related nematodes [Harris TW and Stein LD, 2006]. This database is the central public database to store *C. elegans* research data on classical genetics, cell biology and functional genomics. It includes datasets of phenotype descriptions, RNAi experiments and 3D protein structure, etc. [Schwarz EM et al. 2006, Bieri T et al. 2006]. However, this database is not organized for the storage and manipulation of manually curated gene sets (e.g. genes in the same pathway or that share the same molecular function) [Barrasa MI et al. 2007].

Currently, the Gene Expression Omnibus (GEO) repository [Barrett T et al. 2005],

European Bioinformatics Institute (EBI) ArrayExpress [Parkinson H et al. 2005] and Stanford Microarray Database (SMD) [Demeter J et al. 2007] are the main repositories of large-scale gene expression data in *C. elegans* and include data from spotted microarrays, Affymetrix chips, and SAGE. These sites provide tools for analysis and visualization of gene expression data by publication. However, few tools have been developed for the direct comparison among data sets from different publications.

The databases described above provide storage of comprehensive information but their tools for large-scale dataset analysis target very specific questions. For instance, it is easy to search for all the related information about a certain gene in WormBase. It is convenient to locate and visualize the expression data of one or a few genes from one publication in SMD. Nevertheless, they fail to provide the capability to answer rather basic questions from biologists as follows:
i) Is the experiment result consistent among different publications that studied the same subject (same mutation, same developmental stage, or same pathway, etc.)?
ii) For a particular group of genes, in which experiments do they have strong, correlated gene expression changes?

Aiming at answering such questions, some efforts have been made to conduct generalized and systematical data analysis, including GSEA [Subramanian A et al. 2005], GOTM [Zhang B et al. 2004], and FACT [Kokocinski F et al. 2005]. Gene Set Enrichment Analysis (GSEA) and GOTree Machine (GOTM) derive their power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation [Subramanian A et al. 2005] while Flexible Annotation and Correlation Tool (FACT) allows for detection of important patterns in large data sets by simplifying the integration of heterogeneous data sources and the subsequent application of different algorithms for statistical evaluation or visualization of the annotated data [Kokocinski F et al. 2005].

Several databases have been developed for specialized subjects, such as microarray genes expression data on tumor samples [Stein WD et al., 2007], and EDGEdb: a transcription factor-DNA interaction database focused on *C. elegans* [M Inmaculada Barrasa et al. 2007]. In order to add to these efforts, we developed tools for *C. elegans* gene expression exploration that helps biologists interpret and explore the large-scale gene expression datasets available.

We have created a database consisting of *C. elegans* microarray datasets and SAGE expression data from all available publications from the three repositories (GEO, EBI, SMD) and plus datasets available only from the author's sites. By bringing this data together in one location we facilitate the mining of this gene expression data.

Our database contains 881 arrays from 46 publications covering all the published large scale gene expression studies in *C. elegans*. Moreover, we have developed a

web-based platform, written in Perl CGI, called the Microarray Analysis Display (MAdisplay) for comprehensive and comparative visualization and analysis of gene expression across publications. We provide a full description of the MAdisplay and illustrate its utility through a systematic analysis of the gene expression datasets. We have performed a comprehensive analysis across the whole database to explore the expression patterns of annotated groups of genes in a systematic way.

Chapter 2 Materials and Methods

Data sources

MAdisplay database collects all published *C. elegans* gene expression data. As of May 2007, MAdisplay contains forty-six publications covering many aspects of worm biology (Figure 1a and b). The data were collected from four sources. Most datasets were found in GEO, EBI Array Express, and SMD. SAGE data of three publications: Jones et al. 2001, Halaschek-Wiener et al., 2005, and McKay et al. 2003 is located on the personal site of the publication authors at *C. elegans* Resources (http://elegans.bcgsc.bc.ca/). The distribution of the data that composed our database and their sources are showed in Figure 2.

Data set preprocessing

There are three types of gene expression data in our database: spotted microarrays, Affymetrix GeneChips, and SAGE data. Each type required different preprocessing to produce in a consistent database format. Normalization was also done during preprocessing. For indirect spotted microarray experiments expression is shown relative to the control sample. Affymetrix GeneChip data is $\log_2$ transformed and expression change is shown relative to average of the reference or control samples on a per gene basis. SAGE data was $\log_2$ transformed and then expression values were centered by subtracting the median sample expression from each value.

Moreover, the groups of raw microarray data files that belong to one publication were synthesized into one single dataset file in the common format. A format file is created for each dataset that contains meta information describing the publication and the identifying reference samples. The database also maps gene identifiers used in each experiment to current gene names.

Software

The core of MAdisplay is a Perl CGI program. The input to the program is tab-delimited text files, with a data file and a meta file for each dataset. The user can select one or multiple data sets from the dropdown box, and input or upload a list of

genes of interest. When a request is submitted MAdisplay invokes corresponding format files which direct MAdisplay to open and read data from datasets files. The subset of selected data is passed to the R statistics package for clustering using a custom R clustering function derived from the heatmap.2 function in the gplots library. R is a computer language and open-source software environment for statistical computing and graphics available for most platforms. This R function produces clusters with a consistent grid element size and allows absolute expression value scaling. Output is a synthesized heat map showing the visualized expression level of the target gene(s) in the selected data set(s) represented by values of color along with a tab-delimited text file containing the underlying data.

Gene list analysis

Gene classification is a necessary step before cross-experiment analysis. We included three types of gene lists in our gene library: Publication derived gene lists from figures, tables, and supplemental data in the 46 collected publications; Gene Ontology gene lists in categorized by biological process, cellular component, and molecular function; and KEGG metabolic and signaling pathway genes.

We wrote Perl programs to synthesize all the gene lists in each library with all 127 experiments that derived from the 46 publications (Supplemental Table 1), and created heat maps that using p-value to represent the overall gene expression by certain gene group in certain experiment. R-cluster package was then used to cluster the heat maps and created clustered heat maps, from which we explored the pattern of gene expression associations cross different experiments and gene groups.

Clustering and heat maps

In order to explore the pattern of gene expression across all experiments and all gene lists, we decided to cluster and create heat maps of gene expression across all 127 experiments for each of the five gene list types. Expression values for each gene in the gene lists and each array sample were averaged together to give a single expression value for each gene list and experiment combination. To determine if this gene group expression measure has a particularly large positive or negative value indicating strong expression changes for genes in this group Monte Carlo sampling from the dataset was used. A Holm-Bonferroni corrected p-value for each combination was calculated and the log (p-value) was used to make the heat map. P-values were hierarchically clustered [Eisen MB et al. 1998] by gene list and experiment using the Pearson correlation as the similarity statistic.

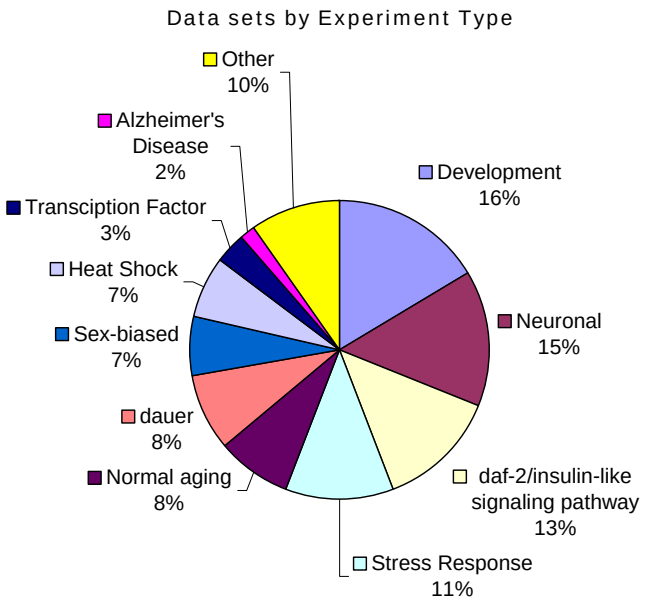


Figure 1a. The distribution of 61 data sets in MAdisplay database by experiment type. The most studied categories are development (10 data sets), neuronal (9 data sets), *daf-2*/insulin-like signaling pathway (8 data sets) and stress response (8 data sets).

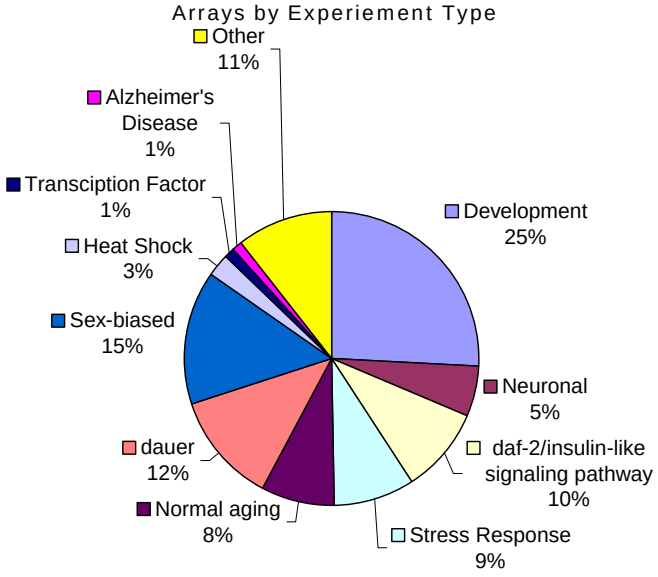


Figure 1b. The distribution of 881 arrays in MAdisplay database by experiment type. The categories with the most experiments are development (228 arrays), sex-biased (130 arrays), dauer stage (106 arrays) and *daf-2*/insulin-like signaling pathway (84 arrays)

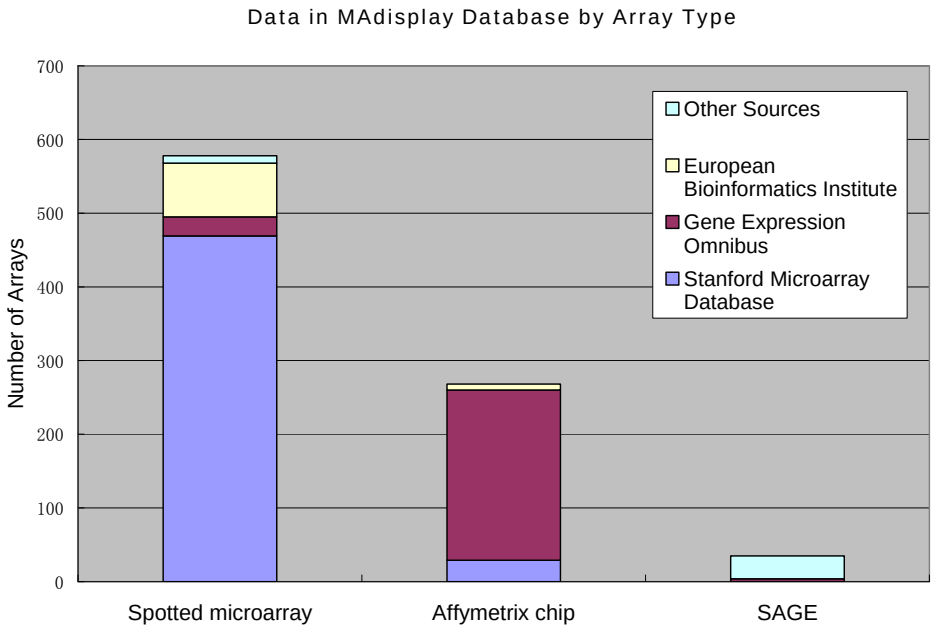


Figure 2. Distribution of samples in MAdisplay database by array type. There are in total 881 samples in our database from the breakdown by source is Stanford Microarray Database (55%), NCBI Gene Expression Omnibus (28%), European Bioinformatics Institute (12%), with the remainder from other sources (4%). Among the 881 samples, 67% are spotted microarrays, 29% are Affymetrix chips, and 4% are SAGE datasets.

Chapter 3 Results

C. elegans microarray database

Experimental data from forty-six publications were collected in our directory-based database via downloading raw data from public data repository websites GEO, EBI and SMD and non-repository datasets. All data are preprocessed to a uniform format that combines separate chip data files that belong to one experiment into one synthesized dataset file. Thus, we produced sixty-one data sets from the forty-six publications. The database includes a total of 881 chip data files and describes 127 different experiments. They are available both as a database for MAdisplay, a freely-accessible web platform that we developed, and as the data bank for the comprehensive analysis we performed.

MAdisplay: a visualization tool

MAdisplay is available to users via the internet directly (http://elegans.uky.edu/gl/madisplay) and through the GeneLists database (http://elegans.uky.edu/gl) via links from each gene list allowing exploration of the genes' expression across different publication data sets.

By inputting genes of interest and selecting datasets from dropdown box, users can view any subset of the data using MAdisplay. (Figure 3a) The result is shown as a clustergram representing expression change by color. Missing data due to single array artifacts or lack of coverage in an experiment is displayed distinctly. The 'download .cdt file' button on the result page allows the user to download the actual data as a tab-delimited text file for subsequent use. Each query generates a unique URL at which the query results are available for two weeks. The 'PubMed' and 'Pub Download' buttons will lead the user to NCBI PubMed entry of the publication and the .pdf file at the publication site respectively (Figure 3b).

MAdisplay merges information from diverse data sources into one comprehensive data set, fulfilling our goal of seeking consistency and facilitating comparison of expression changes from different publications. We next sought to apply functional analysis to identify patterns of correlation between the data sets and the known groups of genes by carrying out a systematic cross-experiment analysis.

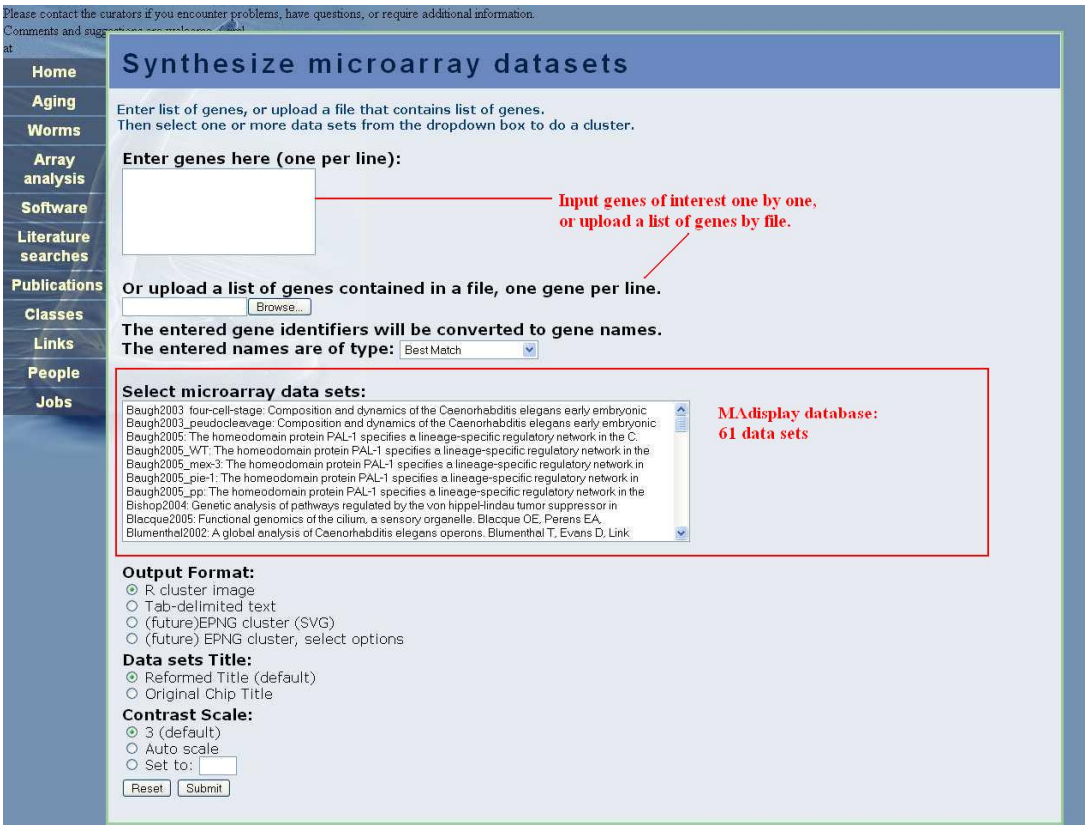


Figure 3a. Screen shot of MAdisplay input interface with annotation (red). Users can input *C. elegans* genes of interest in the textbox or by uploading a list of genes by file. Then the user selects one or multiple data sets, chooses the output format to be either R cluster image or tab-delimited text), the sample annotation format, the cluster image contrast scale, and submits the task.

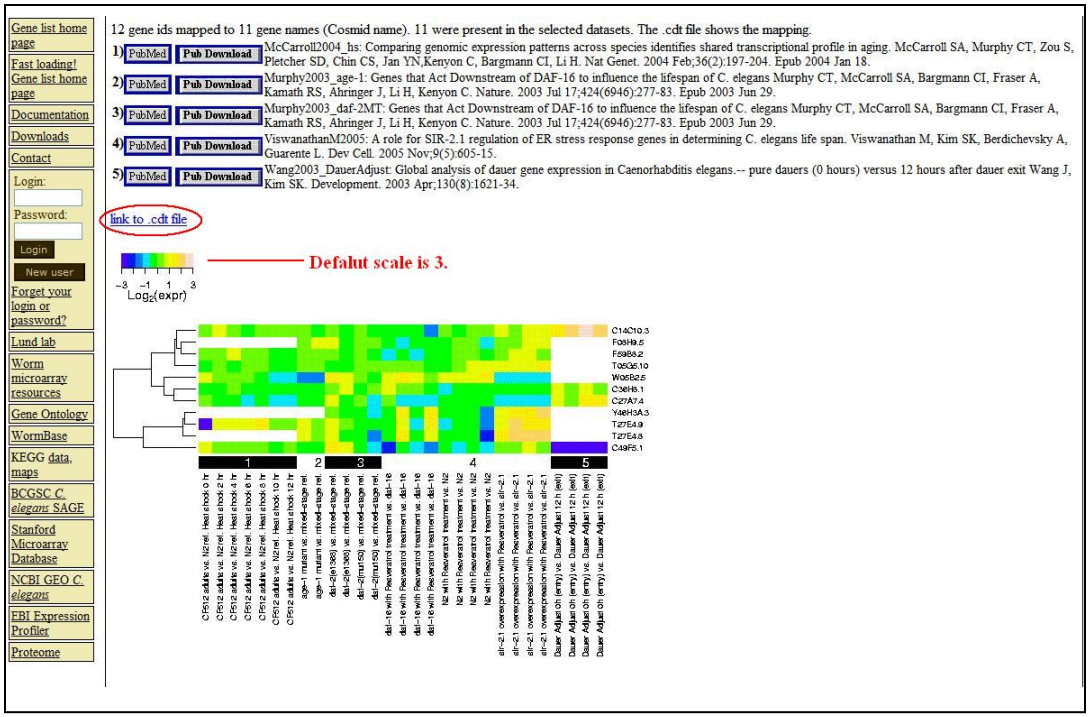


Figure 3b. Screen shot of MAdisplay output interface. The default output type is shown, a clustergram produced by the R cluster package. The default scale runs ±3 in $\log_2$ expression change (8X reduced to 8X increased expression). Curated sample labels are displayed by default in the clustergram and .cdt file. Curated sample labels are based on the labels from original chip and descriptions in the publication and replace the often cryptic original labels. The legend displayed above the clustergram tells the user how many genes of the input list have been found in the selected data sets as well as listing the publications with PubMed and direct journal PDF file download links.

Systematic Cross-experiment Analysis (SCA)

Five categories of gene lists were collected in our gene bank as described in the Materials and Methods: Publication derived gene lists, GO biological processes, GO cellular components, GO molecular function, and KEGG pathway gene lists. We explored the ability of cross-experiment heat maps to provide biologically meaningful insights in the five gene list categories.

In each category, we searched for gene lists whose groupwise expression was significantly altered in any of the experiments. For this purpose, expression values for every gene in the gene lists and each experiment sample were averaged together to give a single expression value for each gene list and experiment combination. To determine if this gene list expression measure has a particularly large value indicating strong expression changes for genes in this group, Monte Carlo sampling from the dataset was used to assign a p-value. A Holm-Bonferroni corrected p-value for each gene list-dataset combination was calculated and the $\log_{10}$ p-values were used to make the heat map. We then used XCLUSTER
[Sherlock G., http://genetics.stanford.edu/~sherlock/cluster.html] using a Pearson correlation to cluster 127 experiments across all gene lists in each of the five gene lists types individually and produced five large clustered p-value heat maps (Supplemental Figures 1-5). The numbers of significant experiments or gene lists in each of the five heat maps are tabulated in Supplemental Table 1.

Some experiments showed no significant expression in one or more of the gene lists categories, but all of them showed significant expression in at least two categories. The numbers of experiments that showed no significant expression in each of the categories are: 2 for publication derived, 2 for GO biological processes, 10 for GO cellular component, 4 for GO molecular function, and 39 for KEGG pathway. Three data sets showed significant expression in only two categories, they are all from Golden's aging study [Golden TR, Melov S. 2004]. This study used a small array with 923 features and thus lack of data is why few gene list enrichments were found.

Publication Gene Lists

We first tested enrichment of the publication derived gene lists. These gene lists are derived from the same publications that provided the datasets in our database. They are a combination of purely data-derived lists and curated lists focused on biologies relevant to the experiments in the publication. Publication gene lists were found to be the best source of gene groupings showing expression changes in multiple experiments. A number of interesting gene list/experiment clusters are found in this heat map.

For example, developmentally regulated genes are up-regulated in response to

xenobiotic compounds and ethanol, under hypoxia, in worms exiting the dauer stage. And the developmentally regulated genes are down-regulated as a function of age, as embryos develop, in response to steroids, and in mutants with under-proliferation of the germline (Figure 4). This result encouraged us to expand our gene list analysis to existing curated gene function databases such as Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) to explore expression patterns of these well studied gene annotation sets.

Publication gene lists are also the best tools to validate the feasibility of our approach. By comparing the heat maps produced by MAdisplay or Systematical Cross-experiment Analysis (SCA) and the original figures or tables in the publications, we should be able to answer questions like:

i) How well did MAdisplay replicate the information in original publication?

ii) How well did Systematic Cross-experiment Analysis represent the information carried by gene groups? Is their any information lost in Systematic Cross-experiment Analysis?

To answer the first question, we randomly picked three gene lists from three different publication and used MAdisplay to profile these gene lists in the publication data sets from which they came. Then, we answered the second question by locating the cross-square of that gene list – data set combination in the Publication Systematic Cross-experiment Analysis heat map. In the first test, a list of 30 genes up-regulated in *vhl-1* null worms compared to wild type was selected from the Bishop *et al.,* 2004 publication. All genes except C31C9.1 showed the same pattern as indicated in the original table, and the cross-square in the SCA heat map is also very significant. In Jones *et al*., 2001, they retested the previous discovery that genes that implicated in longevity showed increased representation in the dauer expression profile and found out that 13 genes are up-regulated in dauer, 6 remains unchanged, 8 are down-regulated. Our result by MAdisplay showed that 11 of those genes' expression increased, 14 remained unchanged or have less than 0.5-fold decrease, 3 decreased more than 1-fold. The result was close to the results in Jones *et al*., but in SCA heat map, the expression pattern was represented as no change as a whole gene group, probably because of the counteraction of increase and decrease within the group.

In our last test, Table 1 of Viswananthan *et al*., 2005 showed that PQN (Prion-like-(Q/N-rich)-domain-bearing) and ABU (Activated in Blocked Unfolded protein response) proteins were induced in wild type by resveratrol, were induced even more in *daf-16* by resveratrol, and were repressed in *sir-2.1* overexpressor worm lines. Our result in both MAdisplay and SCA heat map replicated their conclusion. Based on these sampling tests, we concluded that MAdisplay can replicate most of the information while SCA may lose some power due in gene lists with balanced numbers of up and

down regulated genes. Nevertheless, the Publication gene lists produced the most significant experiments per gene list (52.9) among all five gene lists types which shows its significant utility; and MAdisplay, though more exactly duplicates the original publication data cannot replace the convenience of SCA in analyzing large amounts of data in a systematic manner.

Gene Ontology Gene Lists

The Gene Ontology project, or GO, provides a controlled vocabulary to describe gene and gene product attributes in any organism and covers *C. elegans* with 14,087 gene products annotated and an average of 6.7 annotations per gene. We obtained the gene lists of all three categories in *C. elegans* from the GO website (http://www.geneontology.org/). GO terms with ten or fewer annotated genes were not considered in this analysis. Genes assigned to a GO term are defined to be the genes assigned to the term itself or to its sub-terms.

GO Biological Processes

GO biological processes gene lists contain largest number of lists (771) and produced 13.0 significant experiments per gene list. In the example shown in Figure 5, signaling and regulatory genes specific to metazoan developmental processes are up-regulated in response to xenobiotic compounds and ethonol; in males (compared to hermaphrodites), and in ciliary neurons.

GO Cellular Components

GO cellular components gene lists contain 151 gene lists and produced 12.7 significant experiments per gene list. Our analysis shows that Cytoplasm, Protein complex, and Organelle genes are down-regulated in response to worm treated with xenobiotic compounds and in TGFβ pathway mutants (dauer entry). These gene lists are up-regulated in worms exposed to ethanol, under hypoxia, in the gonad, as embryos develop, and as worms exit the dauer stage (Figure 6a). These are essential genes for organismal growth expressed at high levels during the peak of growth induced in response to some stresses but repressed by xenobiotic exposure.

Intracellular, organelle, and nucleus genes are down-regulated in response to xenobiotic compounds, in TGFβ pathway mutants (dauer entry), males (compared to hermaphrodites), and in germline development mutants while they are up-regulated in worms exposed to ethanol and hypoxia, in gonads, as embryos develop and as worms leave dauer stage (Figure 6b).

GO Molecular Function

GO molecular function gene lists produced 7.9 significant experiments per gene list. We found that Binding and Hydrolase activity genes are down-regulated in worm

somatic tissue (compared to germline) and *daf-c* (dauer formation constitutive) mutants, under nomoxia (compared to hypoxia), in response to xenobiotic exposure (Figure 7a).

Receptor activity and signal transducer activity are up-regulated in response to hormone and xenobiotic compounds; in worm somatic tissue (compared to germline); under nomoxia (compared to hypoxia) (Figure 7b). Xenobiotic exposure triggers a strong and complex biological response. In the worm soma where the neurons are located we find the receptor activity genes highly expressed, a GO term that includes all the neurotransmitter genes, while these genes are poorly expressed in the syncytial gonad.

KEGG Pathway Gene Lists

In order to understand higher-level functions and interactions between the biological system and the natural environment from genomic and molecular information, KEGG metabolic and signaling pathway gene lists were included as our last gene list type in the cross-experiment analysis [Kanehisa M et al 2004 and 2006]. KEGG gene lists turned out to have the fewest number of significant experiments per gene list (3.3). We still found two very significant clusters. In Figure 8a, glycan degradation and urea recycling metabolism genes were strongly up-regulated in early embryos. This may reflect the high rate of protein turnover at the peak of development.

In Figure 8b, four gene lists: Ribosome, Oxidative phosphorylation, ATP synthesis, and Proteasome showed strong regulation in many experiments, increasing as worms develop and under many stress conditions while down-regulated in neurons and under some drug treatments. These groups of genes seem to form a concerted metabolic response to certain events. Neurons surprisingly appear to be a tissue with a low metabolic rate in worms. This is the opposite of what is observed in mammals and perhaps due to the small size of the worm giving very short neuronal processes lengths compared to mammals.

Table 1. Statistics of significant experiments or gene lists in p-value heat maps of the five gene lists types and 127 experiments (p-value < 0.001).

Type of gene lists	# of gene lists	Significant experiments per gene list				Significant gene lists per experiment			
		MIN	MEAN	MEDIAN	MAX	MIN	MEAN	MEDIAN	MAX
Microarray publication derived	295	0	52.9	53	116	0	26.9	14	100
GO biology process	771	0	13	4	92	0	63.9	53	208
GO cellular component	151	0	12.7	2	74	0	15.5	12	65
GO molecular function	259	0	7.9	2.5	77	0	21.7	14	99
KEGG pathway	111	0	3.3	0	73	0	2.7	2	20

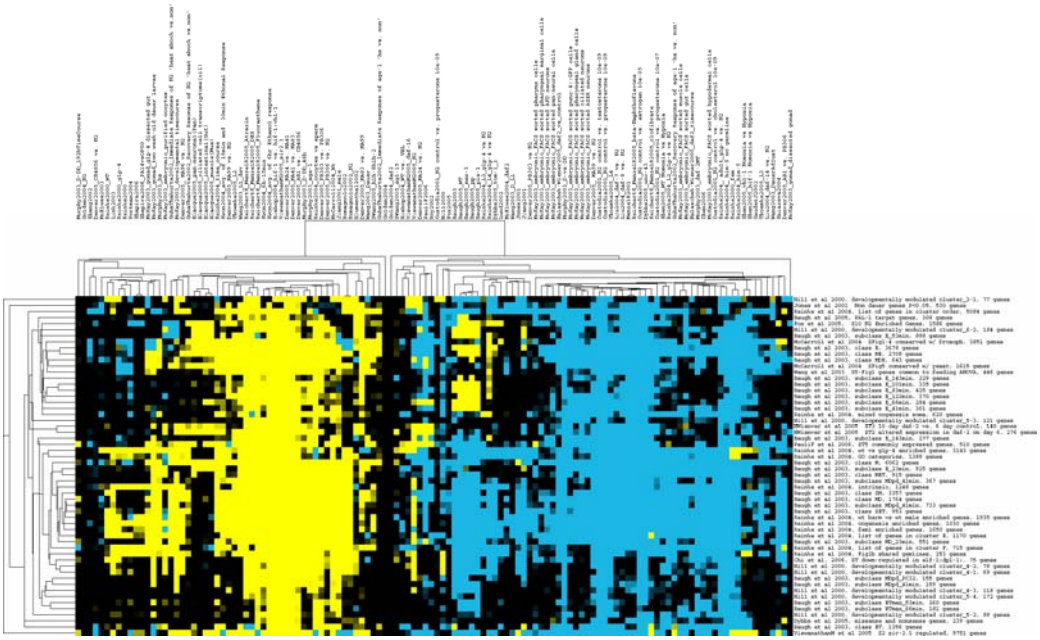


Figure 4. A significant cluster in the Publication gene list heat map.
Developmentally modulated genes are up-regulated in response to xenobiotic compounds and ethanol; under hypoxia; in worms exiting dauer stage; and during oogenesis. These genes are down-regulated as worms age; as embryos develop; in response to steroids; and in under-proliferation germline mutants. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis.

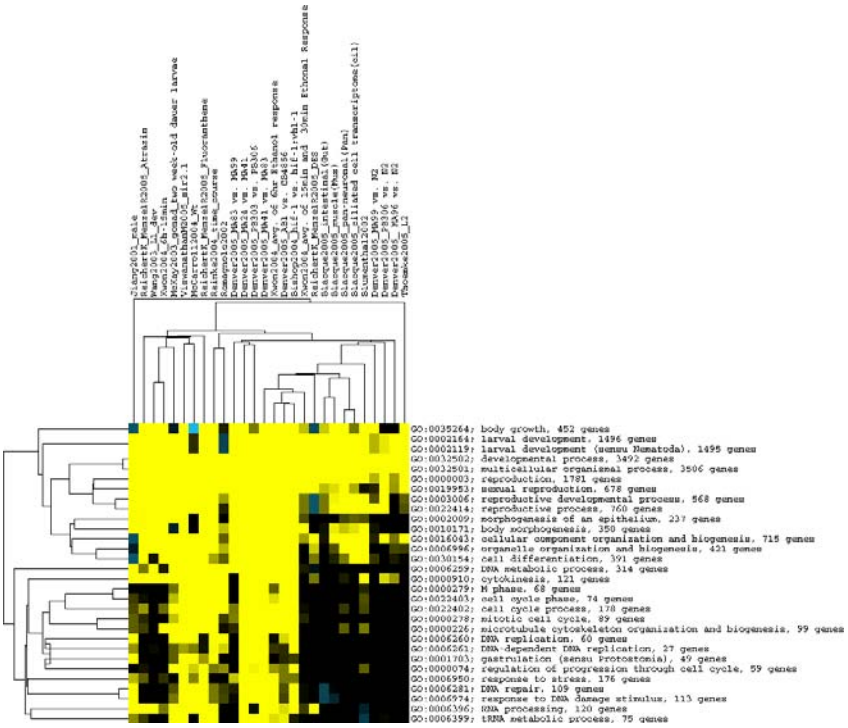


Figure 5. A significant cluster in the GO biology process gene list heat map. Developmental process and multicellular organismal process genes are up-regulated in response to xenobiotic compounds and ethonol; in male (compared to hermaphrodite), and in ciliary neurons. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis.

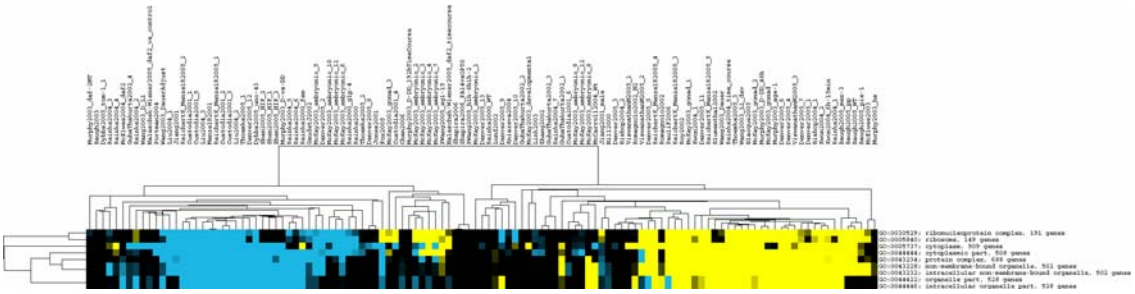


Figure 6a. A significant cluster in the GO cellular components gene list heat map. Cytoplasm, protein complex and organelle genes are down-regulated in response to steroids and xenobiotic compounds; in TGF β signaling pathway mutants. These gene lists are up-regulated in exposure to ethanol and hypoxia; in gonad; at dauer stage. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis.

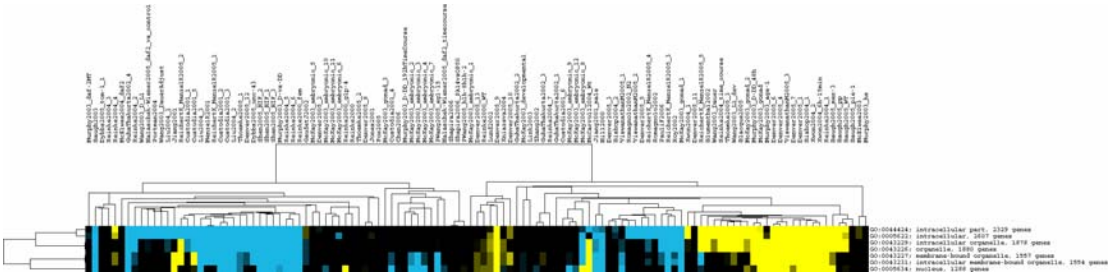


Figure 6b. A significant cluster in the GO cellular components gene list heat map. Intracullular, organelle and nucleus genes are down-regulated in response to steroids; in TGF β signaling pathway mutants, male (compared to hermaphrodite), and mutant worms with abnormal reproduction system while up-regulated in exposure to ethanol and hypoxia; in gonad; at dauer stage. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis.

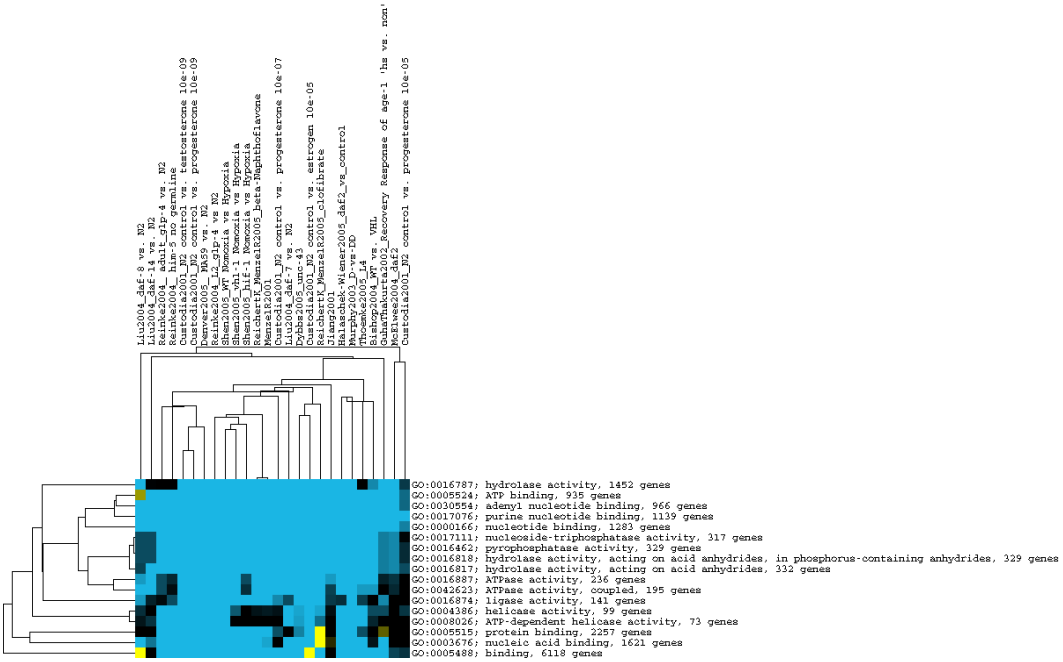


Figure 7a. A significant cluster in the GO molecular function gene list heat map. Nucleotide binding and hydrolase activity genes are down-regulated in worm's soma (compared to germline) and daf-c mutants; under nomoxia (compared to hypoxia); in response to steroids. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis.

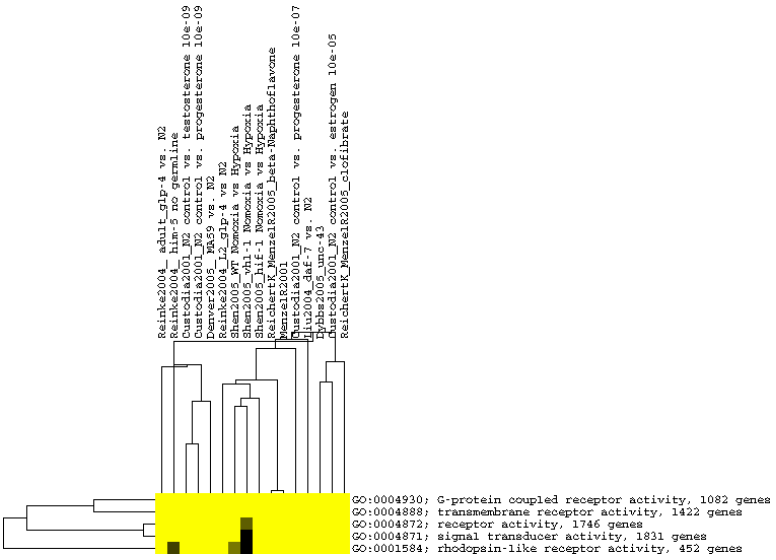


Figure 7b. A significant cluster in the GO molecular function gene list heat map. Receptor activity and signal transducer gene activities are up-regulated in response to steroids and xenobiotic compounds; in worm soma (compared to germline); under nomoxia (compared to hypoxia). A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis.

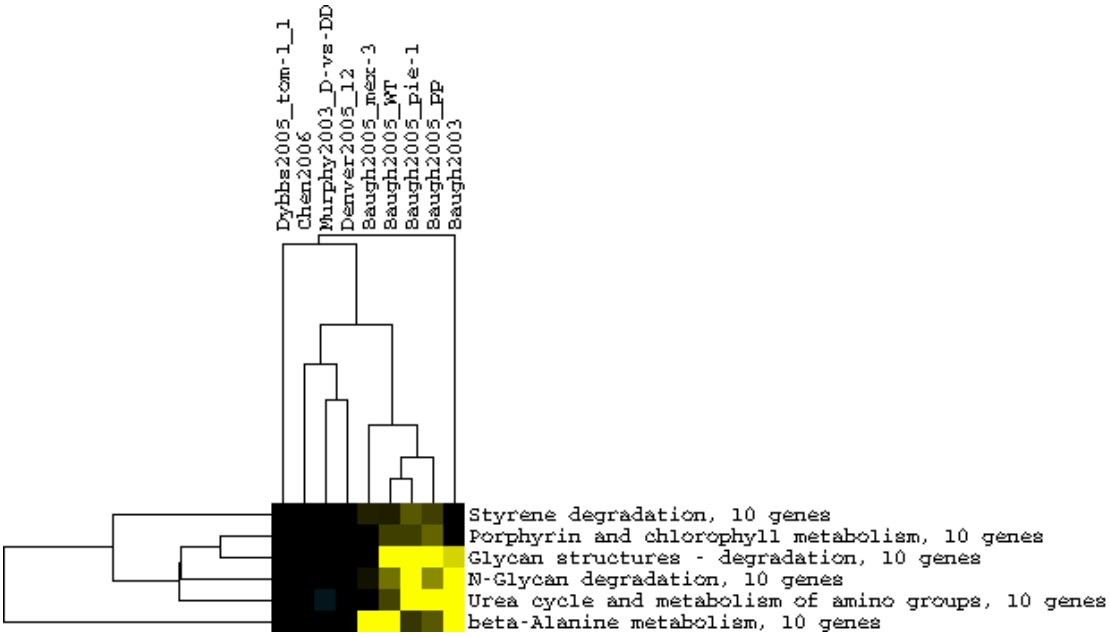


Figure 8a. A significant cluster in the KEGG pathway gene list heat map. Glycan degradation and urea recycling metabolism genes are strongly up-regulated at early embryo stages. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis.

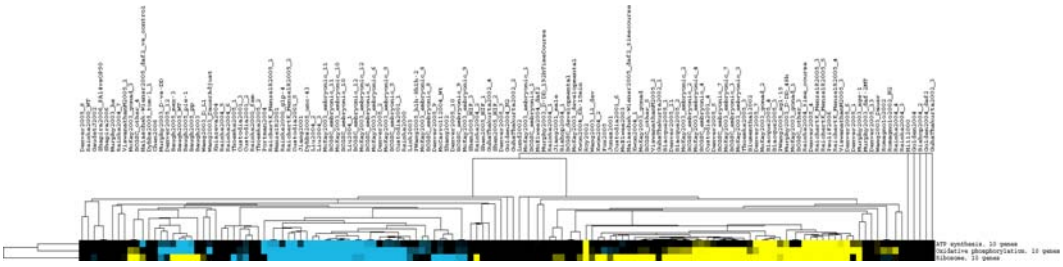


Figure 8b. A significant cluster in the KEGG pathway gene list heat map. Four gene lists: Ribosome, Oxidative phosphorylation, ATP synthesis, and Proteasome showed strong regulation in many experiments, increasing as worms develop and under many stress conditions while down-regulated in neurons and under some drug treatments. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis.

Chapter 4 Discussion

Researchers are now able to generate enormous amount of data in gene expression studies using high-throughput techniques such as microarrays. Most current methods for large-scale data analysis have focused on the analysis of individual genes or individual experiments. Although this is very useful, it is desirable to synthesize experimental results from related experiments to obtain an overall picture of the whole gene expression pattern but in practice this is so difficult it is rarely done. Consequently, new strategies for data mining and organization are needed.

In this study, we have created a database consisting of published gene expression data on *C elegans*. Two things were accomplished with this database. First, we developed a web-accessible program, MAdisplay, which enables biologists to explore and visualize the expression of the genes of interest across selected data sets. The analysis gains power when the expression pattern of selected genes can be retrieved from different publications and displayed together in one clustergram. With the framework in place new datasets can easily be created and added into our database when new studies are published.

Another component of the project, cross-experiment analysis, provided a number of advantages when compared with single experiment analysis. First, it provides a robust way to elucidate a large-scale gene expression patterns by correlating expression changes in one experiment with expression in other experiments in the context of known functional gene groups and pathways.

Second, researchers can focus on gene sets instead of individual genes, which tend to be more interpretable and useful. Third, it can boost the identification of biological responses by combining modest changes in individual genes into functional gene sets that exhibit significant concerted expression changes. Although the risk of the opposite effect, weakening expression when treated as a whole also exists, this effect can be minimized if genes are correctly, that is, biologically, grouped (more specified gene lists, separating known-up-regulated and down-regulated genes, etc).

In addition, the cross-experiment analysis helps link prior knowledge (Gene Ontology, pathway, etc.) to newly published data and itself could be used to refine those manually curated gene sets by uniting or split gene sets according to their expression across our database. Our study concentrates on helping users explore well-organized microarray data to ease the task of data analysis and provides a new perspective on published gene expression information as well.

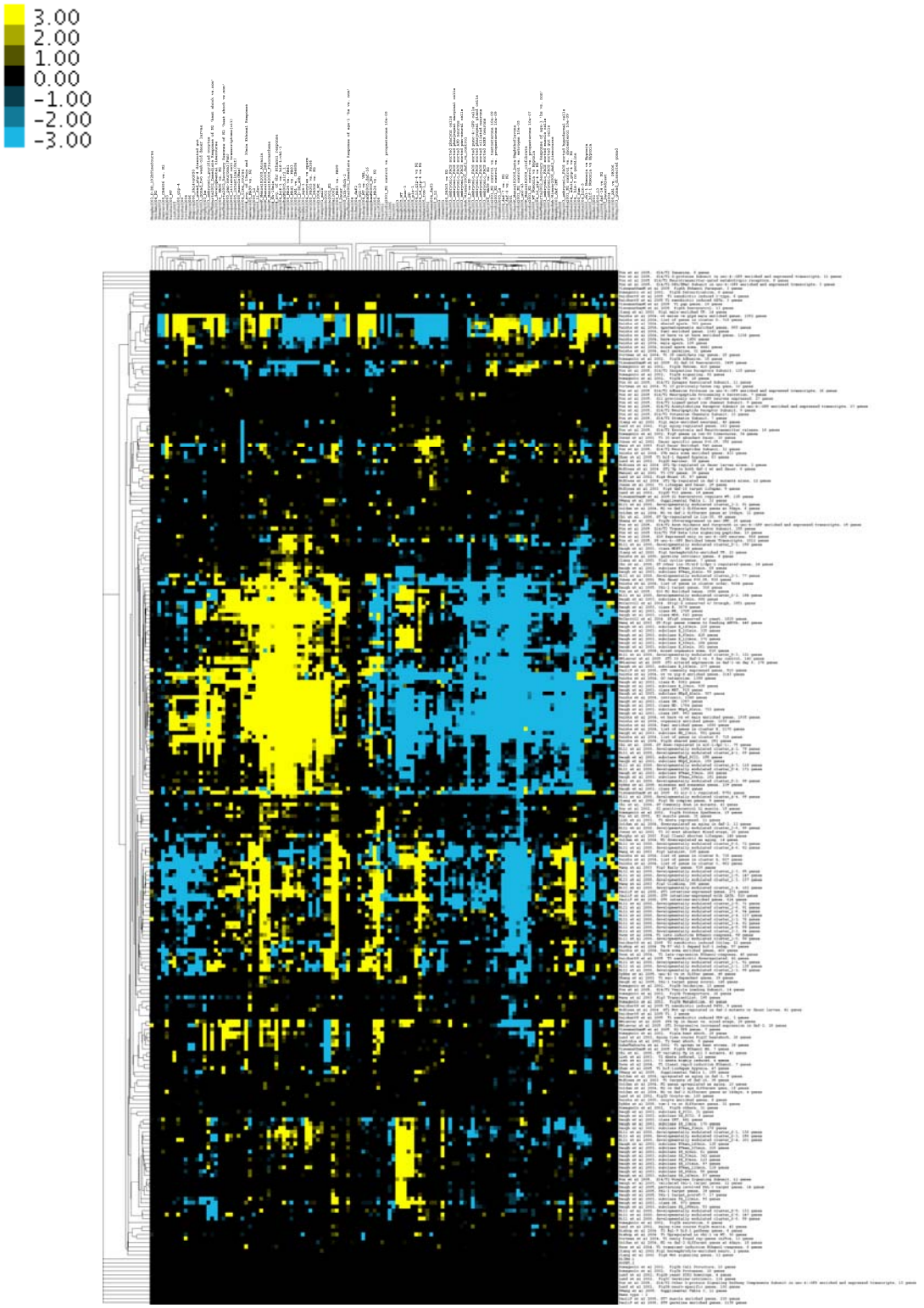
Appendix

Supplemental Table 1. Derivation of 61 data sets, 127 experiments from 46 publications.

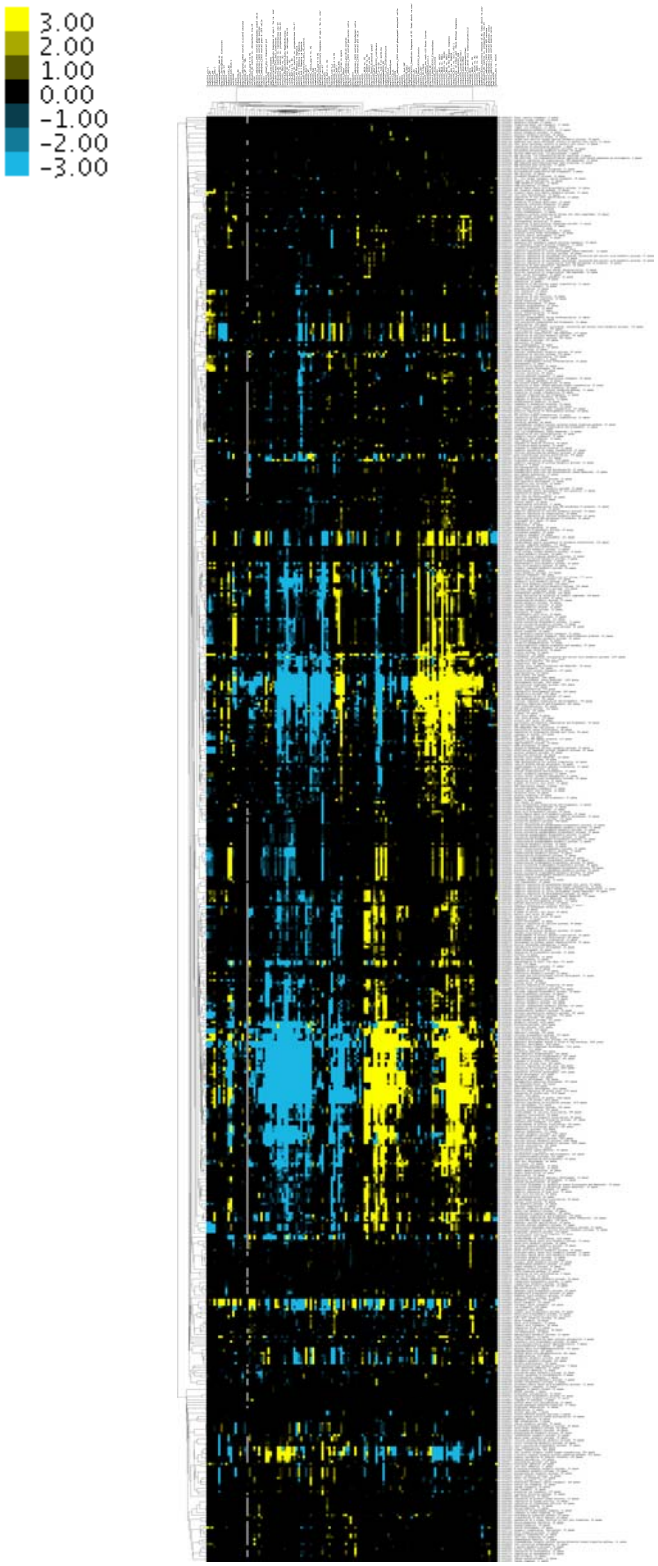
Data Sets	Publication	Array type	Number of arrays	Number of Experiments
1	Baugh et al, 2003	Affymetrix	50	1
2	Baugh et al, 2005	Affymetrix	28	1
3			25	1
4			34	1
5			34	1
6	Bishop et al, 2004	Spotted	4	2
7	Blacque et al, 2005	SAGE	4	4
8	Blumenthal et al, 2002	Spotted	5	1
9	Chen et al. 2006	Affymetrix	4	1
10	Chi et al, 2006	Spotted	10	3
11	Custodia et al, 2001	Spotted	6	6
12	Denver et al 2005	Spotted	40	12
13	Dybbs et al, 2005	Affymetrix	6	1
14			7	1
15	Fox et al, 2005	Affymetrix	7	1
16	Gaudet and Mango, 2002	Spotted	3	1
17	Golden et al, 2004	Spotted	18	1
18			18	1
19	Guha Thakurta et al, 2002	Spotted	3	4
20	Halaschek-Wiener et al, 2005	SAGE	5	1
21				1
22	Hill et al., 2000	Affymetrix	8	1
23	Hristova et al., 2005	Spotted	6	2
24	Jiang et al., 2001	Spotted	29	2
25	Jones et al. 2001	SAGE	2	1
26	Kniazeva et al, 2004	Affymetrix	4	1
27	Kunitomo et al, 2005	Spotted	1	1
28	Kwon et al, 2004	Spotted	7	3
29	Link et al, 2003	Spotted	9	1
30	Liu et al, 2004	spotted	10	3
31	Lund et al, 2002	Spotted	20	1
32	McCarroll et al, 2004	Spotted	7	1
33			8	1

34	McElwee et al, 2003	Spotted	4	1
35	McElwee et al, 2004	Affymetrix	20	1
36	McKay et al, 2003	SAGE	6	1
37			14	13
38			4	3
39	Menzel et al, 2001	Spotted	3	1
40	Murphy et al, 2003	Spotted	3	1
41			30	1
42			4	1
43	Pauli F, et al. 2006	Spotted	8	1
44	Portman et al, 2004	Spotted	7	1
45	Reichert et al, 2005	Spotted	16	5
46	Reinke et al, 2000	Spotted	34	3
47	Reinke et al, 2004	Spotted	76	8
48	Romagnolo et al, 2002	Spotted	20	1
49			20	1
50	Roy et al, 2002	Spotted	6	1
51	Shapira M et al, 2006	spotted	18	2
52	Shen et al, 2005	Affymetrix	29	3
53	Szewczyk NJ et al, 2006	Spotted	3	1
54	Thoemke et al, 2005	Spotted	10	3
55	Viswanathan M, 2005	Spotted	12	3
56	Wang J et al, 2003	Spotted	42	2
57			8	1
58			44	1
59	Wang P et al, 2005	Affymetrix	6	1
60			6	1
61	Zhang et al 2002	Spotted	6	1
61 Data Sets	**46 publications**		**881 arrays**	**127 experiments**

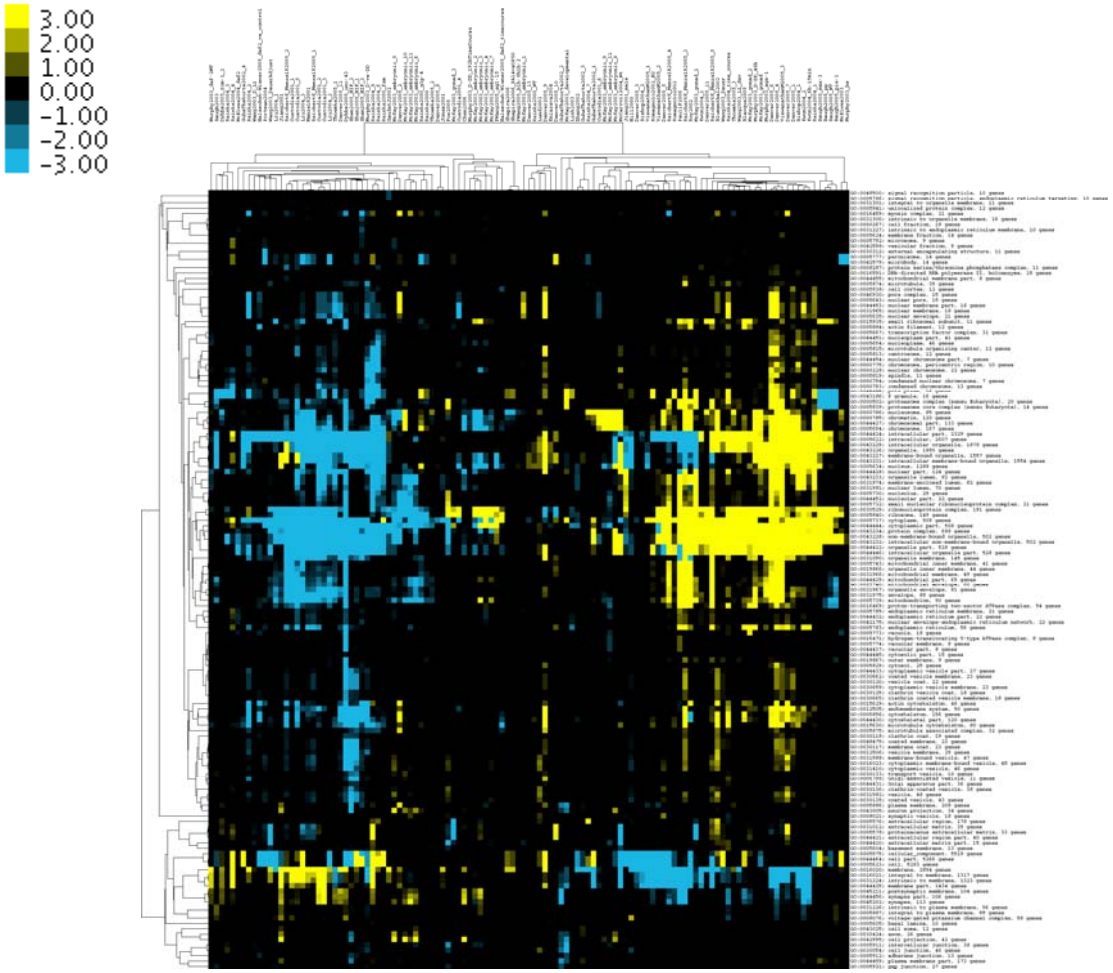
Supplemental Figure 1. Heat map of Publication Gene Lists across 127 experiments.
A larger version of this figure is available at
http://elegans.uky.edu/Xue_thesis/.



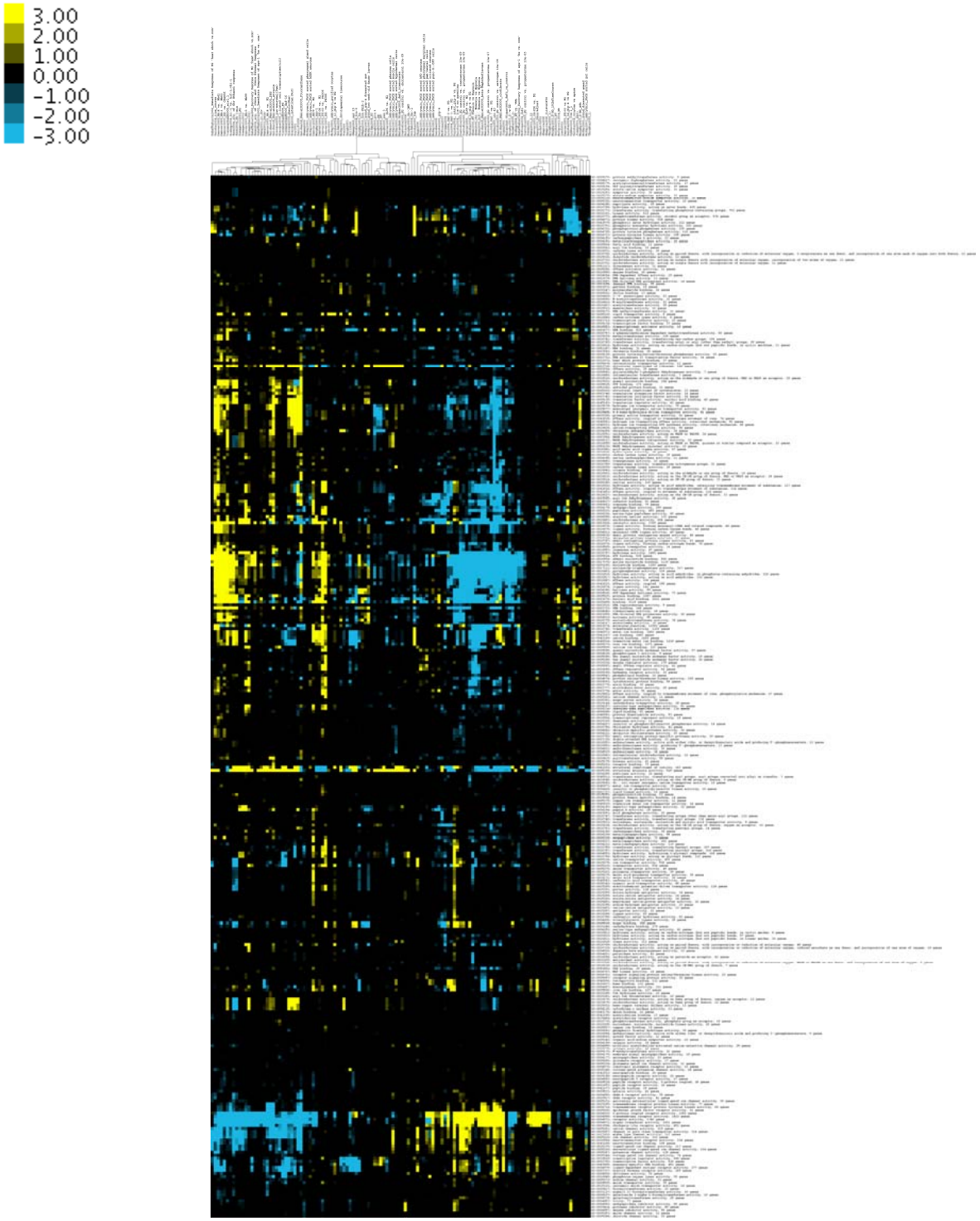
Supplemental Figure 2. Heat map of GO Biological Processes Gene Lists across 127 experiments. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis/.



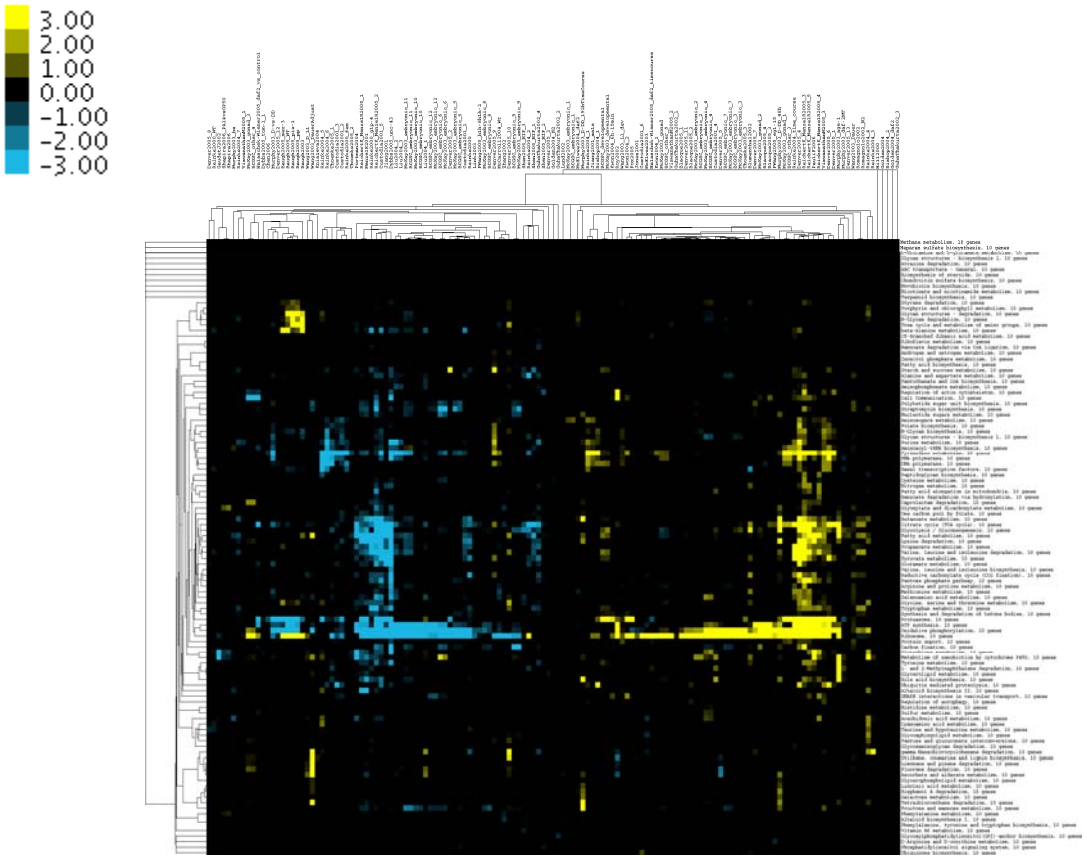
Supplemental Figure 3. Heat map of GO Cellular Components Gene Lists across 127 experiments. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis/.



Supplemental Figure 4. Heat map of GO Molecular Function Gene Lists across 127 experiments. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis/.



Supplemental Figure 5. Heat map of KEGG Gene Lists across 127 experiments. A larger version of this figure is available at http://elegans.uky.edu/Xue_thesis/.



References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: The genome sequence of Drosophila melanogaster. Science. 2000 Mar 24; 287(5461):2185-95.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J. Mol. Biol. (1990) 215:403-410. Medline

Astin J, Merry A, Rajan J, Kuwabara PE: Caenorhabditis elegans functional genomics: Omic resonance. Brief Funct Genomic Proteomic. 2004 Apr;3(1):26-34

Barrasa MI, Vaglio P, Cavasino F, Jacotot L, Walhout AJ: EDGEdb: a transcription factor-DNA Interaction database for the analysis of C. elegans differential gene expression. BMC Genomics. 2007; 8: 21.

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE,

Fujibuchi W, Edgar R. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. 2005;33:D562–6. doi: 10.1093/nar/gki022.

Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP: Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development. 2003 Mar;130(5):889-900

Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, Chan J, Chen N, Chen WJ, Davis P, Fiedler TJ, Girard L, Han M, Harris TW, Kishore R, Lee R, McKay S, Müller HM, Nakamura C, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Spooner W, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Durbin R, Stein LD, Sternberg PW, Spieth J: WormBase: new content and better access. Nucleic Acids Res. 2007 Jan;35(Database issue):D506-10. Epub 2006 Nov 11.

Bishop T, Lau KW, Epstein AC, Kim SK, Jiang M, O'Rourke D, Pugh CW, Gleadle JM, Taylor MS, Hodgkin J, Ratcliffe PJ: Genetic analysis of pathways regulated by the von Hippel-Lindau tumor suppressor in Caenorhabditis elegans. PLoS Biol. 2004 Oct;2(10):e289. Epub 2004 Sep 7.

Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G, Ball CA: The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res 2007 Jan 1;35(Database Issue):D766-770

Drmanac, S, Stavropoulos NA, Labat I, Vonau J, Hauser B, Soares MB, and Drmanac R: Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics, 37:29–40, 1996.

Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998, 95:14863-14868.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: Life with 6000 Genes. Science 25 October 1996: Vol. 274. no. 5287, pp. 546 - 567 DOI: 10.1126/science.274.5287.546

Golden TR, Melov S. Microarray analysis of gene expression with age in individual nematodes. Aging Cell. 2004 Jun;3(3):111-24

Gollub J and Sherlockv G: Clustering Microarray Data. METHODS IN ENZYMOLOGY (2006), VOL. 411.

Han, J., and Kamber, M. (2001). Data Mining: Concepts and Techniques. (Morgan Kaufmann, San Francisco, CA), Ch 8.

Harris TW, Stein LD: WormBase: methods for data mining and comparative genomics. Methods Mol Biol. 2006;351:31-50

Hill AA, Hunter CP, Tsung BT, Tucker-Kellogg G, Brown EL: Genomic analysis of gene expression in C. elegans. Science. 2000 Oct 27;290(5492):809-12.

Jones SJ, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R, Marra MA : Changes in gene expression associated with developmental arrest and longevity in Caenorhabditis elegans. Genome Res. 2001 Aug;11(8):1346-52.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D277-80.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D354-7.

Kokocinski F, Delhomme N, Wrobel G, Hummerich L, Toedt G, Lichter P: FACT – a framework for the functional interpretation of high-throughput experiments. BMC Bioinformatics 2005, 6:161 doi:10.1186/1471-2105-6-161

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, and Brown EL: Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol.,14: 1675–1680, 1996

McKay SJ, Johnsen R, Khattra J, Asano J, Baillie DL, Chan S, Dube N, Fang L, Goszczynski B, Ha E, Halfnight E, Hollebakken R, Huang P, Hung K, Jensen V, Jones SJ, Kai H, Li D, Mah A, Marra M, McGhee J, Newbury R, Pouzyrev A, Riddle DL, Sonnhammer E, Tian H, Tu D, Tyson JR, Vatcher G, Warner A, Wong K, Zhao Z, Moerman DG: Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. Cold Spring Harb Symp Quant Biol. 2003;68:159-69.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999 Jan 1;27(1):29-34.

Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S, Brazma A: ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2005;33:D553–5. doi: 10.1093/nar/gki056.

Pleasance ED, Marra MA and Jones SJ: 'Assessment of SAGE in transcript identification', Genome Res.2003; Vol. 13, pp.1203–1215.

Reinke V, Smith HE, Nance J, Wang J, Van Doren C, Begley R, Jones SJ, Davis EB, Scherer S, Ward S, Kim SK: A global profile of germline gene expression in C. elegans. Mol Cell. 2000 Sep;6(3):605-16

Roy PJ, Stuart JM, Lund J, Kim SK: Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. Nature. 2002 Aug 29;418(6901):975-9.

Schwarz EM, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Canaran P, Chan J, Chen N, Chen WJ, Davis P, Fiedler TJ, Girard L, Harris TW, Kenny EE, Kishore R, Lawson D, Lee R, Muller HM, Nakamura C, Ozersky P, Petcherski A, Rogers A, Spooner W, Tuli MA, Van Auken K, Wang D, Durbin R, Spieth J, Stein LD, Sternberg PW: WormBase: better software, richer content. Nucleic Acids Res. 2006;34:D475–8. doi: 10.1093/nar/gkj061.

Stein WD, Litman T, Fojo T, Bates SE: A database study that identifies genes whose expression correlates, negatively or positively, with 5-year survival of cancer patients. Biochimica et Biophysica Acta 1770 (2007) 857–871

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS published online Sep 30, 2005; doi:10.1073/pnas.0506580102

Sulston JE, Horvitz HR: Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev Biol. 1977 Mar;56(1):110-56.

Sulston JE, Schierenberg E, White JG, Thomson JN : The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol. 1983 Nov;100(1):64-119.

The *C. elegans* Sequencing Consortium: Genome Sequence of the Nematode *C. elegans*:

A Platform for Investigating Biology. Science 11 December 1998: Vol. 282. no. 5396, pp. 2012 - 2018 DOI: 10.1126/science.282.5396.2012

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. (2000) 25: 25-29

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995;270:484–487.

Viswanathan M, Kim SK, Berdichevsky A, Guarente L: A role for SIR-2.1 regulation of ER stress response genes in determining C. elegans life span. Dev Cell. 2005 Nov;9(5):605-15.

Werner T. Regulatory networks: linking microarray data to systems biology. Mech Ageing Dev. 2007 Jan;128(1):168-72. Epub 2006 Nov 20

Zhang B, Schmoyer D, Kirov S, Snoddy J.: GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 2004, 5:16

Vita

Author's Name: Lin Xue

Birthplace: Shanghai, China

Birthdate: July 27, 1982

Education
Bachelor of Science in Life Sciences
Fudan University
June 2004

Research Experience
University of Kentucky
Lexington, KY
Aug. 2004 – present
Graduate Research Assistant / Graduate Teaching Assistant

Fudan University
Shanghai, China
Oct. 2002 – June 2004

Honors, Awards, and Activities
- Gennexttech Young Investigators Scholarship, 2006
- GenNext Technologies Summer Training, Aug. 2006
- University of Kentucky Graduate School Academic Year Fellowship, 2006
- University of Kentucky Teaching Assistant Scholarship, 2004-2005, 2007
- Fudan University People's Fellowship, 2001-2004
- Shanghai Excellent Student Award, 2000

Abstracts
Xue, L, Lund J. Tools for microarray data analysis and their application to a cross-experiment analysis of gene expression in *C. elegans*. 16th International *C. elegans* Meeting. UCLA, 2007.

Xue, L, Lund J. Cross-experiment gene expression analysis in *C. elegans*. UT-ORNL-KBRIN Bioinformatics Summit, 2007.

Xue, L, Lund J. A visualization tool to synthesize gene expression datasets of *C. elegans*. UT-ORNL-KBRIN Bioinformatics Summit, 2006.

Xue, L, Lund J. A visualization tool to synthesize gene expression datasets of *C. elegans*. Tri-State Worm Meeting 2006.