

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

The statistical and molecular logic of gene expression patterns in *Caenorhabditis elegans*

Permalink

<https://escholarship.org/uc/item/9hd8g5g5>

Author

McCarroll, Steven Andrew

Publication Date

2005

Peer reviewed|Thesis/dissertation

The statistical and molecular logic
of gene expression patterns in *Caenorhabditis elegans*

by

Steven Andrew McCarroll

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Neuroscience

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



Acknowledgements

My advisor, Cori Bargmann, was a source of inspiration, thoughtful guidance, and encouragement. Her support of my graduate research throughout its various incarnations helped me to develop in new directions as a scientist. Her lab was a wonderful environment in which to pursue graduate research.

Hao Li was a patient teacher and a stimulating collaborator who became a second mentor to my graduate work.

Jesse Gray, Chris Patil, and Coleen Murphy were inspired sources of friendship and collaboration.

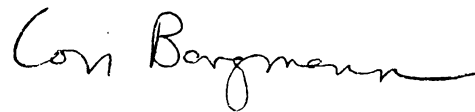
Annika Malmberg made these years a delight.

**The statistical and molecular logic
of gene expression patterns in *Caenorhabditis elegans***

Steven Andrew McCarroll

Abstract

Gene regulation uses transcriptional control systems with a molecular logic we seek to understand. Genome-scale sequence and expression data increasingly make it possible to use genomic patterns in sequences and gene expression levels to reveal the logic of transcriptional regulation. In this dissertation, two approaches to understanding transcriptional regulation are developed and applied. First, we describe a novel method for identifying phylogenetic conservation in genomic transcriptional patterns. We use this new approach to identify gene expression programs in aging, development, and mRNA degradation that are shared by organisms as diverse as the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, the yeast *Saccharomyces cerevisiae*, and the human *Homo sapiens*. We use this approach to search databases of gene expression patterns to identify relationships among the physiological programs of diverse organisms. Second, we use a statistical approach, probabilistic segmentation, to identify candidate transcriptional control sequences in the promoters of a large gene family, the chemosensory receptor genes in *C. elegans*. We identify many new candidate transcriptional control sequences and show that one of these is a novel E-box motif that confers expression in the ADL chemosensory neurons.



ncis

RY

UNIVERSITY

UNIVERSIT

LIB

San

L

UNIVERSITY

ncis

RY

UNIVERSITY

UNIVERSIT

LIB

San

L

UNIVERSITY

UNIVERSIT

ncis

RY

UNIVERSITY

UNIVERSIT

LIB

San

Table of contents

Introduction	1
Chapter 1. Comparative functional genomics: Using conservation to understand gene expression patterns	5
Chapter 2. Coming of age in <i>metazoa</i> : What young adulthood can tell us about aging, lifespan, metabolism, and disease	63
Chapter 3. Decoding the transcriptional regulation of chemosensory receptor genes in <i>C. elegans</i>	87
Appendix. A useful modular system for making <i>C. elegans</i> expression constructs	121

cis
RY
VERSITY
VERSIT.
LIB
m
L
IV
cis
RY
UNIVERSIT
UNIVERSIT
LIB
m
L
ALIFORN
ALIFORN
nci
AR
UNIVERSIT
UNIVERS
LIB

List of Tables

Chapter 1: Comparative functional genomics

Table 1: Closest *C. elegans* matches to *D. melanogaster* aging 32

Table 2: Cross-species searches of DNA microarray databases 33

Chapter 2: Coming of age in *metazoa*

Table 1: Conservation of MYA between humans and invertebrates 82

Chapter 3: Decoding the transcriptional regulation of chemoreceptor genes

Table 1: Distribution of word lengths and occurrences 111

Table 2: Positional preference of sequence motifs 112

Table 3: Functional preference of sequence motifs 113

Table 4: Sequence context of chemoreceptor E-box sites 114

Table 5: Expression patterns of chemoreceptor genes with E-boxes 115

Table 6: Alignment of DNA-binding domains of bHLH proteins 116

ncis
RY
UNIVERSITY
UNIVERSIT
181
an
L
UNIVERSIT
ncis
RY
UNIVERSIT
UNIVERSIT
181
an
L
ALFORN
ALFORN
nci
AR
UNIVERSIT
UNIVERS
181
an

List of Figures

Chapter 1: Comparative functional genomics

Figure 1: Comparative functional genomic analysis schematic	35
Figure 2: Correlated regulation of orthologous genes by aging	37
Figure 3: Temporal distribution of conserved gene regulation	39
Figure 4: Aging, lifespan, and conserved early-adult change	41
Suppl. Figure 1: Conserved transcriptional signature in aging	43
Suppl. Figure 2: Various conserved transcriptional signatures in aging	45
Suppl. Figure 3: Shared transcriptional signature in larval development	51
Suppl. Figure 3: Shared transcriptional signature in embryo development	54
Suppl. Figure 3: Shared transcriptional signature in gametogenesis	58

Chapter 2: Coming of age in *metazoa*

Figure 1: Adult-onset gene expression program shared by humans, flies	83
Figure 2: Conserved metazoan adult-onset gene expression change	84
Figure 3: Timing of repression of oxidative metabolism genes	85
Figure 4: A model for the regulation and effects of MYA	86

Chapter 3: Decoding the transcriptional regulation of chemoreceptor genes

Figure 1: Positional preference of candidate regulatory motifs	117
Figure 2: Distribution of motif across chemoreceptor subfamilies	120

icis

RY

UNIVERSITY

UNIVERSIT

LIB

an

L

UNIVERSITY

icis

RY

UNIVERSIT

UNIVERSIT

LIB

an

L

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

INTRODUCTION

The advent of genome-scale sequence and expression data make it possible to study gene regulation on a large scale. Perhaps more importantly, it also allows researchers to ask entirely new kinds of questions about transcriptional systems, enabling researchers to go beyond questions about individual genes to begin to scrutinize the meaning of global *patterns* of gene expression. This development, while exciting, requires much: biology needs to develop new approaches, statistical tools, even a vocabulary for understanding genomic patterns and patterned change.

In chapter 1, we describe a novel method for inferring the meaning of patterns of gene expression, by comparing these patterns to patterns from other organisms. The principle of homology across organisms – the idea that the most important features of a gene, protein, or biological process are shared by different organisms – has been profoundly useful in biology. We sought to discover whether such homologies could be identified at the level of genome-scale patterns of gene expression, and if so, whether such conservation could be used to discover, validate, and draw insight from analogies among the physiological programs of different organisms. In Chapter 1, we find such

cis

RY

VERSITY

IVERSI

IBI

an

L

ITY

cis

RY

IVERSIT

IVERSI

IBI

an

L

ALIFORN

ALIFORN

anci

ARC

IVERSI

IVERSI

IBI

an



homologous patterns of change in gene expression for a wide range of biological processes, including adult aging, development, and mRNA degradation. We also develop an approach for using these kinds of homologies to search databases of gene expression patterns, much as BLAST is now used to search databases of gene and protein sequences. We find that such an approach can be used to identify novel, suggestive homologies among the physiological programs of different organisms.

One of the most surprising findings in Chapter 1 is the discovery of homologous patterns of gene expression change in adult aging. Chapter 2 generalizes and further explores the biological significance of these results. The conserved patterns of gene expression change observed earlier in *Caenorhabditis elegans* and *Drosophila melanogaster* turn out also to be present in *Homo sapiens*. Moreover, the early-adult timing of these shared transcriptional changes is also conserved in rodents. This is the first finding of a conserved, metazoan young adult gene expression program, and we discuss its potential relationship to current questions and controversies in aging, lifespan, and metabolism research. We also propose ways in which this adult-onset gene expression program may contribute to changes in the pattern of vulnerabilities to disease that humans experience as they progress through adulthood.

The transcriptional regulation of a gene is determined by the non-protein-coding sequences around and within it, most frequently by its upstream, promoter sequence. While scientists have long understood how the protein-coding parts of genes encode protein sequences, comparatively little is understood about how regulatory sequence determines the place, time, and manner of gene expression. The advent of genome-scale sequence data for many

ETS

RY

ERSITY

IVERSI

BI

an

L

IV

ncis

RO

UNIVERSIT

IVERSI

BI

an

L

ALFORN

ALFORN

ncis

ARC

UNIVERSI

UNIVERS

IBI

an

organisms offers new opportunities for decoding regulatory sequences. In particular, the sequences of large families of functionally related genes, which are likely to share many regulatory motifs, may offer new opportunities for finding such transcriptional control sequences. In Chapter 3, we use a statistical approach, probabilistic segmentation, to identify many candidate transcriptional control elements in the promoters of the *C. elegans* chemosensory receptor gene family. Many of the motifs we find show positional preference, are specific to chemosensory receptor genes, and correspond to the binding sites of known families of transcription factors in different organisms. We functionally characterize one of these motifs, the E-box sequence WWYCACSTGY, and find that it confers expression in the ADL chemosensory neurons.

The research described in this dissertation is also a kind of meta-experiment: how can the analysis of statistical patterns in genome-scale biological data sets contribute to our understanding of biological problems? Such approaches, if successful, have the potential to contribute novel kinds of inference and to result in extremely rapid progress in fields; and if unsuccessful, can result in much pointless data-gazing. A primary theme that emerges is the extent to which these approaches identify a biological object (a gene, a gene expression program, or a biological sequence) much more readily than they identify the function of that object. This represents an inversion of classical approaches in biology, which begin with a function (such as a mutant phenotype, or a protein activity) and then strive toward identifying an underlying biological object (such as a specific lesioned gene, or a purified protein). For example, in chapter 1 we identify homologous adult-onset gene expression programs in highly diverged metazoans. The existence of such broad,

18

Y

ERSIT

VERSI

BI

an

L

THE

18

an

L

anc

AR

UNIVERSIT

UNIVERSI

18

an

L

ALFORN

ALFORN

18

an

L

anc

AR

UNIVERSI

UNIVERS

18

an

shared patterns of adult-onset transcriptional change in worms, flies, and mammals is undoubtedly biologically important; but clarifying *for what* it is important requires somewhat more inference (Chapter 2), and we cannot yet manipulate this gene expression program to assess its function directly. In Chapter 3, we identify a set of short sequences in the *C. elegans* chemoreceptor gene family whose statistical signatures imply that they are transcriptional control elements. We characterize one of these sequences functionally, finding that drives expression in particular chemosensory neurons, but the functional characterization is a late rather than initial step in this investigative process. In both studies, much progress is derived from the existence of large, computable data sets on gene function and protein localization, particularly from the Gene Ontology (GO) annotation system. Such data sets allow the investigator to connect emerging biological patterns to the great body of work and insight that has come from classical approaches.

Data-driven approaches like those described here may increasingly supplement the function-driven approaches that we already wield as experimental biologists. If so, then the style of investigation in this dissertation may be a taste of things to come.

UCSF LIBRARY

cis

RY

ERSIT

ERSIT

IBI

an

2

LIB

ci

RO

IVERSIT

IVERSIT

IBI

an

2

ALFORN

ALFOR

nci

ARC

UNIVERSI

UNIVERSI

IBI

an

2

CHAPTER 1

COMPARATIVE FUNCTIONAL GENOMICS: USING CONSERVATION TO UNDERSTAND GENE EXPRESSION PATTERNS

Abstract

We describe a method for systematically comparing gene expression patterns across organisms, by using genome-wide comparative analysis of DNA microarray experiments. Analogous gene expression programs are identified, comprising shared patterns of regulation across orthologous genes. Biological features of these patterns are identified as highly conserved subpatterns that correspond to Gene Ontology categories. We demonstrate these methods by analyzing a specific biological process, aging, and then show that similar analysis can be applied to a range of biological processes. We find that two highly diverged metazoans, the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*, implement a shared adult-onset gene expression program of genes involved in mitochondrial metabolism, DNA repair, catabolism, peptidolysis, and cellular transport. Using this approach to search databases of gene expression data, we find conserved transcriptional signatures in larval development, embryogenesis, gametogenesis, and mRNA degradation.

Introduction

Gene expression profiling measures the expression levels of thousands of genes at once (DeRisi et al., 1997; Lipshutz et al., 1999). While most expression profiling studies have focused on the specific genes that respond to specific conditions, another important direction in functional genomics is to derive insight from global patterns of gene expression. Genome-scale expression

patterns have been used as physiological fingerprints for classifying tumors(Chung et al., 2002; Ramaswamy et al., 2003) and assigning uncharacterized mutations and drugs to known pathways(Hughes et al., 2000). Because they use information from many genes at once, patterns have great discriminating power, even when the transcriptional effects on individual genes are small(Hughes et al., 2000; Mootha et al., 2003).

The patterns of gene expression changes observed in microarray experiments can be extensive and complex. To try to analyze these patterns, we have exploited the principle that important biological processes are often conserved between organisms. We present an approach to comparative functional genomics based on shared patterns of regulation across orthologous genes. We also present a method for identifying conserved biological components of those patterns that correspond to Gene Ontology categories. These methods can be used to search databases of microarray experiments to discover connections among biological processes in different organisms.

Results

We used phylogenetic analysis to systematically identify orthologous groups of genes for all pairwise comparisons between *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Homo sapiens* (**Supplementary Tables 1-5**). For *C. elegans* and *Drosophila* we identified 3,851 most-conserved orthologous gene pairs (**Fig. 1a**).

DNA microarrays were used in each organism to compare gene expression under different conditions (**Fig. 1b**). The gene phylogenetic relationships were then used to systematically match measurements of differential expression

UCSF LIBRARY

between orthologous genes from the two organisms (**Fig. 1c**). The correlation of the log-fold-change in expression of orthologous genes was used to assess the extent of shared regulation (**Methods**).

Using this approach, we asked whether gene expression patterns in adult aging were shared by two highly diverged metazoans: the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*, whose last common ancestor existed about one billion years ago (Wang et al., 1999). We used spotted-PCR-product microarrays (DeRisi et al., 1997) to compare gene expression in middle-aged adult (6 days adult age) and young adult (0 days adult age) sterile *C. elegans* hermaphrodites, and used Affymetrix oligonucleotide microarrays (Lipshutz et al., 1999) to compare expression in middle-aged adult (d23) and young adult (d3) female flies (Pletcher et al., 2002). The cross-species Pearson correlation of the log-change in expression of orthologous during aging was 0.144, which is significant at the 10^{-11} level. 16 comparisons of independent experimental replicates were all highly significant, with a mean correlation of 0.155 ± 0.012 ($p < 10^{-35}$). These results indicate that most aging-related changes are species-specific; nonetheless, the conserved component of these expression profiles could include several hundred *C. elegans*-*Drosophila* ortholog pairs. This result is highly statistically significant; it is not observed in one million randomized pairings of the expression results (**Fig. 2a**). Non-parametric tests confirmed the statistical significance of the shared regulation (Spearman rank correlation: 0.156, $p < 10^{-12}$; Kendall's Tau: 0.106, $p < 10^{-12}$).

Similarly correlated regulation during aging was observed in microarray data sets from different tissues, labs, and experimental platforms. We used Affymetrix microarrays to compare gene expression in heads of young and aged

181

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

VERSIT

adult male flies, and observed a similar correlation with aging *C. elegans* (0.148 ($p < 10^{-11}$)). These results suggest that the conserved regulation is present in *Drosophila* somatic tissue. A published profile of adult aging in *C. elegans* using Affymetrix microarrays (Hill et al., 2000) also showed highly significant correlations when compared to profiles of aging from *Drosophila* heads ($R=0.180$, $p < 10^{-6}$) and profiles of aging whole female *Drosophila* ($R=0.150$, $p < 10^{-6}$). Both genes that have orthologs in the yeast *Saccharomyces cerevisiae*, and genes that have no homology to any gene in yeast, showed highly significant correlations in the regulation of orthologs by aging (Fig. 2A).

The statistical and explanatory power of gene expression analysis is greatly increased by grouping related genes into functional categories. The Gene Ontology (GO) annotation system (Ashburner et al., 2000) was used to define hundreds of groups of ortholog pairs with common molecular function, cellular localization, or biological role. We searched the data sets for highly conserved subpatterns that corresponded to GO categories (see Methods).

Fourteen GO categories showed highly conserved patterns of regulation in aging *Drosophila* heads and aging *C. elegans* (Fig. 2b, Supplementary Fig. 1), at a strict significance cutoff at which less than one false positive category would be expected by chance. No categories showed significant negative correlation. Similar comparisons using other published aging data sets (Hill et al., 2000; Pletcher et al., 2002) and other time points yielded a broadly overlapping set of GO categories (Supplementary Fig. 2), confirming the robustness of the result.

Both aging *Drosophila* heads and aging *C. elegans* repressed genes in GO categories for mitochondrial membrane and mitochondrial inner membrane (Fig. 2b), including many components of the mitochondrial respiratory chain, the ATP

synthase complex, and the citric acid cycle. Earlier studies have identified individual oxidative metabolism genes that are repressed by aging in worms, flies, or mammals (Kayo et al., 2001; Lee et al., 1999; Pletcher et al., 2002); our results suggest that these individual results are manifestations of a broad, conserved pattern that includes most oxidative metabolism genes. *C. elegans* and *Drosophila* also showed conserved patterns of regulation of genes for peptidases, catabolism, and DNA repair proteins (Fig. 2b).

An unexpected shared feature of aging in *C. elegans* and *Drosophila* was the repression of orthologous genes involved in diverse ATP-utilizing molecular transport functions, including primary active transporters, ion transporters, and ABC transporters (Fig. 2b). Aging appears to involve a decreased transcriptional commitment to active intracellular and intercellular movement of ions, nutrients, and transmitters.

Most transcriptional changes were specific to worms or to flies. For example, in these experiments and work by Lund *et al.* (Lund et al., 2002), *C. elegans* aging repressed collagen genes and induced genes that encode histones, transposases, and DNA and RNA helicases; these changes did not characterize *Drosophila* aging. *Drosophila* induced many genes encoding cytochrome p450s, glycosylases, and peptidoglycan receptors, but *C. elegans* did not alter the expression of the orthologous genes.

Two specific molecular features of aging -- the repression of oxidative metabolism genes (Kayo et al., 2001; Zou et al., 2000) and the correlation between transcriptional profiles of aging and stress (Zou et al., 2000) -- are widely assumed to represent responses to oxidative damage with advancing age. By profiling gene expression at time intervals throughout adulthood in *C. elegans* and

VERSIT
VERSI
18
an
L
VERSIT
VERSIT
18
an
L
ALFORN
ALFORN
ancl
AR
NIVERS
NIVERS
18
an

Drosophila, we assessed how conserved gene expression programs were implemented over time. To our surprise, both the conserved global pattern of change in gene expression (**Fig. 3a, 3b**) and the conserved repression of oxidative metabolism genes (**Fig. 3c**) were abruptly implemented early in adulthood. We profiled the transcriptional responses of worms and flies to heat and oxidative stress, and found that stress responses are significantly correlated with early-adulthood transcriptional programs in both organisms (**Fig. 3d**). These results suggest that changes in gene expression with adult age are not solely implemented in response to cumulative damage. Instead, the timing of these conserved features of aging suggests developmentally timed transcriptional regulation in young adults.

To increase the power and generality of comparative analysis, we developed methods for searching databases of gene expression profiles from different organisms, much as BLAST allows researchers to find related gene and protein sequences in different species. We assembled, from our own experiments and 300 published *C. elegans* experiments (Hill et al., 2000; Sherlock et al., 2001), a database of *C. elegans* expression profiles addressing larval development, sex differences, aging, environmental stress responses, neuronal signaling, organogenesis, dauer formation, and developmental defects (**Supplementary Table 4**). We then queried this database with the *Drosophila* aging data, by ranking the *C. elegans* expression profiles in this database according to their similarity to profiles of *Drosophila* aging (**Methods**).

Strikingly, the *C. elegans* profiles most similar to search profiles of *Drosophila* aging were profiles of *C. elegans* aging (**Table 1**). This cross-species similarity persisted across data sets from different *C. elegans* labs, *Drosophila* labs,

specific experimental designs, and microarray platforms (Tables 1, 2). The next closest *C. elegans* matches to the *Drosophila* aging profiles were profiles of heat stress responses. Aging and heat stress are related in *C. elegans*: many long-lived mutants are thermotolerant, and mild heat stress increases longevity (Finkel and Holbrook, 2000; Lithgow et al., 1995). The strongest negative correlation with expression profiles of *Drosophila* aging came from profiles of *daf-2(e1368)* mutants (Murphy et al., 2003). The *daf-2* gene encodes an insulin/IGF-1 receptor homolog; *daf-2* mutants age more slowly and live twice as long as wild-type animals (Kenyon et al., 1993; Kimura et al., 1997). *Drosophila* gene expression profiles thus appear to identify both analogous and related gene expression experiments in *C. elegans*.

To extend database searching to other biological questions, we searched the database of *C. elegans* gene expression profiles using published profiles of *Drosophila* larval development (Arbeitman et al., 2002). Among all *C. elegans* expression profiles, the closest matches to profiles of *Drosophila* larval development were profiles of *C. elegans* larval development (Jiang et al., 2001) (Table 2). Shared patterns of regulation were observed across GO categories for protein processing, protein transport, secretion, and macromolecule catabolism (Supplementary Fig. 3).

Published profiles of embryonic development in *Drosophila* (Arbeitman et al., 2002) were used to search the *C. elegans* gene expression experiments. The best matches were comparisons of gene expression in *C. elegans* embryos to expression in larvae (Hill et al., 2000; Jiang et al., 2001), and comparisons of embryonic expression in different mutants (Gaudet and Mango, 2002) (Table 2). Shared patterns of change included GO categories for cell cycle, DNA

metabolism, cytoskeleton, microtubule-based processes, and proteolysis (Supplementary Fig. 4).

To assess whether database searching could make connections among more diverged organisms, we searched the *C. elegans* database with expression profiles of sporulation in the yeast *Saccharomyces cerevisiae* (Chu et al., 1998). The strongest matches to profiles of yeast sporulation came from profiles of germ line formation in *C. elegans* (Reinke et al., 2000) (Table 2). The database match appeared to recognize conserved transcriptional programs associated with meiosis; significant GO categories in the match between yeast sporulation and nematode germ line development included categories for nucleoplasm, chromosome condensation, and DNA strand elongation (Supplementary Fig. 5).

The Stanford Microarray Database (Sherlock et al., 2001) contains 647 publicly available *S. cerevisiae* experiments and 2247 *H. sapiens* experiments. We generated a table of ortholog pairs in yeast and man to allow searches between these databases (Supplementary Tables 5, 6). Human mRNA degradation has been profiled in T-cells by blocking transcription with actinomycin D, then using microarrays to measure subsequent transcript abundance (Raghavan et al., 2002). The strongest matches to this array experiment among all yeast experiments were profiles comparing *rpb1*, the RNA polymerase II mutant, to wild-type yeast (Wang et al., 2002b) (Table 2). Since both experiments represent a transcriptional block, the similarity of these profiles suggests that mRNA stability is conserved for orthologous genes in yeast and man. GO categories for kinases and transcription factors were among the most rapidly degraded mRNAs in both humans and yeast, while transcripts encoding ribosomal and core metabolic proteins were extremely stable in both organisms. Searching the

CU

U

VERSIT

VERSI

BI

U

U

U

nci

R

IVERSIT

IVERSI

BI

U

U

IFORM

IFORM

nci

R

NIVERS

NIVERS

BI

U

human database with the yeast *rpb1* profiles yielded experiments that may correspond to transcriptional blockade: profiles of host responses to diverse pathogenic infections (Boldrick et al., 2002; Cuadras et al., 2002; Detweiler et al., 2001; Guillemin et al., 2002) and profiles from whole blood, which is dominated by mRNAs from erythrocytes, which lack nuclei and therefore do not perform transcription (Whitney et al., 2003).

Discussion

We have developed a way to discover analogies among biological processes in diverse organisms by comparative analysis of gene expression patterns. These methods are freely available on the paper's accompanying web site.

We used this approach to identify a shared pattern of adult-onset gene regulation that is implemented by two highly diverged metazoans, *C. elegans* and *Drosophila*. An unexpected feature of this conserved program was the repression of genes encoding orthologous transporter-ATPases, which offers a candidate mechanistic connection between two known features of aging: reduction in ATP synthesis, and decline in the physiological activity of neurons, muscle, and excretory processes. An expected feature of this conserved program was the repression of genes with roles in mitochondrial oxidative respiration. Surprisingly, however, we found that worms and flies both repressed these genes early in adulthood, before the onset of functional decline, and more abruptly than a damage-response model would predict.

In *C. elegans*, mitochondrial respiration before early adulthood limits subsequent adult lifespan, but later mitochondrial respiration is not lifespan-

ci
R
VERSIT
VERSI
81
m
L
nci
R
UNIVERSIT
UNIVERSI
81
m
L
UNIVERSIT
UNIVERSIT
nci
AR
UNIVERSIT
UNIVERSIT
81
m

limiting (Fig. 4)(Dillin et al., 2002b). At about the same time that this transition takes place, the insulin pathway begins to regulate lifespan(Dillin et al., 2002a). Mammals also begin to lose oxidative capacity early in adulthood(Somani et al., 1992), and certain longevity-limiting effects of the insulin pathway on fat accumulation begin early in adulthood(Blucher et al., 2003). Our results show that the transformation of these relationships early in adulthood is accompanied by a conserved transcriptional program. An exciting direction in aging research will be to identify the signals that induce conserved physiological change early in adulthood.

While these results suggest the potential of systematic comparative analysis in functional genomics, we expect that future work will improve upon our methods. For example, the development of ways to systematically assign genes to regulons(Segal et al., 2003; Wang et al., 2002a) may make possible regulon-based measures of correlation that could be more sensitive and specific in their identification of analogous biological programs. The integrative use of expression data from different species represents an emerging area of research(Alter et al., 2003; Stuart et al., 2003; Teichmann and Babu, 2002; van Noort et al., 2003; Whitfield et al., 2002), and elements of these different approaches might be combined to develop additional tools. Our computational approach is also readily generalized to data on protein expression and modification(Gygi et al., 1999).

Comparative functional genomics could be a powerful way to distinguish the essential from the species-specific features of biological processes such as disease, stress, and development. Aided by growing repositories for expression data(Brazma et al., 2003; Edgar et al., 2002) and conventions for reporting

genomic experiments(Stoeckert et al., 2002), measures of correlation in searchable databases could identify novel analogies among disease states, mutant strains, and drug responses in diverse organisms.

Methods

Phylogenetic analysis. Sequence data were obtained from Gadfly Release 2 of the Berkeley *Drosophila* Genome Project and Wormpep version 51, from the Sanger Center. These protein sets were merged and subjected to all-against-all BLASTP analysis using the BLOSUM62 substitution matrix, DUSTSEQ complexity filtering, and a probability cutoff of 10^{-10} . The BLAST results were used to group *C. elegans* and *Drosophila* genes into clusters by means of an agglomerative clustering algorithm described elsewhere (Rubin et al., 2000). Agglomerative clustering yielded 5,042 clusters of 2 to 161 genes each.

Multiple sequence alignment was performed for each of the 5,042 clusters using CLUSTALW with default parameters. The sequence alignment for each cluster was used to generate a phylogenogram by the neighbor-joining method, also using CLUSTALW. Points at which the resulting phylogenograms branched into species-specific clades defined orthologous groups. For *C. elegans* and *Drosophila*, 3,851 groups were thus defined. If an orthologous group contained more than one gene for either species (1,290 cases, the result of additional branching after species divergence), the most-conserved orthologous gene pair was identified by comparing pairwise Smith-Waterman alignment scores. In the resulting ortholog table, each orthologous group was thus represented by a single pair of genes. An alternative method of identifying ortholog pairs, by identifying all mutual best hits directly from the BLAST results, yielded an ortholog table 90% identical to that yielded above, without significantly changing any of the subsequent statistical results. This approach was used to build ortholog tables for each pairing of *C. elegans*, *D. melanogaster*, *H. sapiens*, and *S. cerevisiae*.

1815

1815

UNIVERSITY

UNIVERSITY

UNIVERSITY

1815

1815

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

Strains, lifespans, and culture conditions. All *C. elegans* experiments were performed on a CF512 *fer-15(b26)* II; *fem-1(hc17)* IV mutant strain, whose spermatids fail to activate into spermatozoa at 25°C. Culturing this strain at 25°C prevented self-fertilization and therefore eliminated contributions from embryonic transcripts. Several *C. elegans* mutants fail to develop a germ line; however, given the important role that the germ line plays in regulating aging and life span (Hsin and Kenyon, 1999), such strains do not offer a way to profile normal adult aging. The CF512 strain has normal germ line stem cells and oocytes, ages at the wild-type rate, and has the same life span as wild-type N2 animals. Adult age in *C. elegans* is measured from the first day of adulthood, after adult anatomy, adult behaviors, and reproductive maturity are established. *C. elegans* has a median adult life span of about ten days at 25°C, with fecundity that peaks on the second day of adulthood and is largely exhausted by the fourth day. To yield synchronized *C. elegans* populations for analysis, we axenized eggs and then synchronized animals via L1 arrest, as described elsewhere (Lewis and Fleming, 1995). For the aging experiments, samples were collected at 0, 8, 16, 28, 40, 52, 70, 96, and 144 hours (6 days) after animals reached adulthood.

Experiments with tissue from whole *Drosophila* have been described elsewhere (Pletcher et al., 2002). These experiments used the Dahomey strain, an outbred stock whose life span is similar to that of newly caught, wild populations. In *Drosophila*, adult age is measured from eclosion, when fully-formed adults emerge from the pupal case. The median Dahomey female adult life span at 25°C is 28 days; fecundity peaks at d10 and is nearly exhausted by d21.



For experiments with tissue from *Drosophila* heads, the w1118 strain was cultured in standard cornmeal agar medium. The w1118 strain is an inbred lab stock widely used in genetic and transgenic studies. w1118 males have a median life span of 35 days. Adult males were collected within 24 hr after eclosion. Approximately 200 flies were maintained in constant darkness in each food bottle at 25°C and 70% humidity and were transferred to fresh bottles every 3-4 days. Transcripts were harvested at d3 and d47.

***C. elegans* expression profiling.** For *C. elegans*, 18,455 predicted *C. elegans* genes were amplified by PCR using oligos obtained from Research Genetics. The sequences of these oligos have been deposited in Wormbase. These PCR products were printed on glass slides using techniques described elsewhere (Murphy et al., 2003). The microarrays were used to survey gene expression by comparing mRNAs extracted from each time point sample to a common mixed reference mRNA pool via competitive hybridization. RNA was extracted with Trizol and labeled according to standard techniques.

***Drosophila* expression profiling.** The experiments with tissue from whole *Drosophila* have been described elsewhere (Pletcher et al., 2002). At least 4 replicate Affymetrix roDROMEGA GeneChips were hybridized for each sample point; these replicates were derived from independent RNA extractions of separate biological samples. The results for replicate GeneChips were consistent, with correlations (of log-fold-expression measurements) all exceeding 0.90.

Samples from *Drosophila* heads were processed and analyzed in a different laboratory from the whole-*Drosophila* experiments. To separate the head from the rest of the body, flies were frozen and briefly vortexed in liquid nitrogen. Fly heads were collected using a sieve which retained fly bodies. Total RNA was

extracted using Trizol (GIBCO/BRL). Poly(A) RNA was isolated using Oligotex resin (Qiagen). Samples were profiled with Affymetrix DrosGenome1 GeneChips, using standard Affymetrix protocol.

Expression profiles of heat and oxidative stress. To profile the effect of heat stress on *C. elegans* gene expression, CF512 animals were synchronized and cultured as described above. CF512 adults were then exposed to 30°C (experimental condition) or maintained at 25°C (control condition) for 2, 4, 6, 8, 10, and 12 hours. Corresponding experimental and control samples were compared by competitive hybridization to DNA microarrays as described above.

To profile the effect of oxidative stress on gene expression in *Drosophila*, w118 male adult flies were cultured as described above. At adult d2, the flies were fed sucrose with 15mM paraquat (experimental condition) or regular sucrose (control condition) for 30 hours. Heads were collected and transcripts extracted as described above, then profiled using Affymetrix GeneChips.

Microarray data processing. Except where individual experimental replicates are discussed in the text, we generally used a composite gene expression profile that represented the average of experimental replicates. To construct such composite profiles, log-fold-change measurements were averaged across replicates for each probe. Profiles of differential gene expression (comparing two different samples or conditions from the same organism) were obtained in the following ways. In experiments in which two-channel microarrays were used to compare two experimental samples directly, those measurements of log-fold-change were used directly. In experiments in which multiple two-channel microarrays were used to compare multiple experimental samples to a common reference sample, the log-fold-change measurements from

different hybridizations were differenced to compare the corresponding experimental samples, removing the effect of the reference sample. In experiments in which multiple single-channel Affymetrix microarrays were used to profile multiple experimental samples, normalized log-fold-expression measurements were differenced to compare experimental samples.

Calculation of interspecies correlations. The Pearson correlation (r) of the log-fold-change measurements for orthologous genes was used to measure global correlation between heterologous expression profiles ($r = \frac{\sum_{i=1..n} (x_i - \mu_x)(y_i - \mu_y)}{n \sigma_x \sigma_y}$, where $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ are vectors of log-fold-induction measurements for orthologous genes in *C. elegans* and *Drosophila*, and μ and σ are the mean and standard deviation of these measurements). Statistical significance of Pearson correlations was assessed using Student's t test ($t = r \sqrt{(n-2)/(1-r^2)}$) with $n-2$ degrees of freedom, where n is the number of ortholog pairs yielding gene induction measurements in both organisms.

Monte Carlo simulations. 2040 ortholog pairs yielded expression measurements in both organisms at both the old and young time points, allowing measurements of log-fold-induction with age. In each simulation, we randomly paired these 2040 *C. elegans* and 2040 *Drosophila* genes, then calculated the Pearson correlation of their respective experimental log-fold-change measurements. Across one million such simulations, the Pearson correlation was distributed in accordance with Student's t distribution. The distribution of simulated correlations had a mean of zero, a standard deviation of 0.022, and the following percentiles: 95th 0.037, 99th 0.053. The largest observation was 0.094.

Assessment of potential artifacts. Artifacts in the profiling process can introduce subtle trends which, if common to both profiles, could cause artifactual measured correlations. Although the cross-platform nature of interspecies comparisons makes such shared trends much less likely, an artifact of potential concern involves a potential relationship between measurements of differential hybridization and a gene's overall hybridization strength (a function of transcript abundance and GC content, both of which are correlated for orthologous genes). To assess the potential contribution of such effects, we repeated the Monte Carlo simulation, but rather than pairing genes randomly, we paired genes which were in the same quantile for overall hybridization intensity. The resulting distribution of correlations did not show a positive bias or a significantly greater variance.

Non-parametric statistical tests. For each Pearson correlation presented in this paper, the Spearman rank correlation and Kendall's Tau were also calculated; these three assessments of significance for global correlations were in broad agreement for all of the results discussed here.

Gene Ontology analysis. The Gene Ontology system organizes biological processes, biochemical functions, and cellular compartments ("terms") on a directed graph that describes the relationships among these terms (Ashburner et al., 2000). Each term on the GO graph defines a subgraph which consists of the term, its more-specific sub-terms, and the genes associated with those terms. For example, the subgraph for the term "ion channel" includes genes associated with the "potassium channel" and "voltage-gated ion channel" terms. For each GO term and its associated subgraph, the contribution of associated ortholog pairs to the global Pearson correlation was measured by the

partial summation of the Pearson correlation $r_j = \sum_{i \in J} (x_i - \mu_x)(y_i - \mu_y) / n \sum_{i \in J} \sum_{i \in J}$ (where J is the set of ortholog pairs associated with the GO category), using the global mean and variance from the entire gene induction profiles. The distribution of r_j is well approximated by a normal distribution with zero mean and standard deviation $n_j^{1/2}/n$, where n_j is the number of ortholog pairs in J . The significance of r_j was assessed using the z test with $z = r_j / (n_j^{1/2}/n)$. Only subgraphs with expression data for a useful number (10-100) of gene pairs were analyzed; there were about 250 such subgraphs for *C. elegans* – *Drosophila* comparisons, depending on the particular experiments compared. **Fig. 2B** and **Supplementary Fig. 1-5** directly represent the results of this analysis for different microarray data sets, showing expression data for the ortholog pairs in those GO categories that had significant z -scores.

Statistical controls for Gene Ontology analysis. To bootstrap the false positive rate for these multiple, non-independent hypothesis tests, we repeatedly shuffled the expression data and re-performed the analysis 10,000 times. Applying a test statistic cutoff of 3.0 in a two-sided test, the estimated false positive rate (average number of GO categories with $|z| > 3.0$) from the randomized data was 0.73 ± 0.62 , consistent with the false positive rate of 0.65 expected for the z test. To assess whether conserved gene regulation was significantly *concentrated* into Gene Ontology categories, versus being randomly distributed across the genome, we performed the following additional control: starting with the correlated experimental data sets, we randomized the assignment of paired measurements to ortholog pairs, then re-performed the Gene Ontology analysis. The positive rate (average number of GO categories with $|z| > 3.0$) was 1.40 ± 0.91 . By contrast, the actual data sets had fourteen

significant GO categories, a result that was not obtained in 10,000 simulations.

Analogous results were obtained for the results in Supplementary Fig. 1-5.

Databases of microarray data. We downloaded all publicly available *C. elegans*, *S. cerevisiae*, and *H. sapiens* microarray data from the Stanford Microarray Database (SMD)(Sherlock et al., 2001). We used the pixel regression correlation (cutoff=0.6) to filter individual gene measurements, then obtained the log-ratio-of-medians for each probe for each experiment. We used only those profiles for which the identity of the original experiment was provided; this gave us about 300 *C. elegans*, 650 *S. cerevisiae*, and 2,250 *H. sapiens* gene expression profiles. For the *C. elegans* database, we added our own aging and heat stress experiments (another 40 profiles) and carefully subjected all profiles to cross-replicate averaging and cross-reference differencing, as described above under microarray data processing. To search a database using a gene expression profile from one organism as a query, we ranked all the profiles in the database by their similarity to the query profile, using the similarity metric described above. In Table 2, we present the three closest matches from each database search.

Accession numbers. Microarray data sets have been deposited in NCBI's Gene Expression Omnibus(Edgar et al., 2002) and assigned the following accession numbers: GSE832, GSM12883, GSM12884, GSM12885, GSM12886, GSM12887, GSM12888, GSM12889; GSE 826, GSE827, GSM 12770, GSM12772, GSM12773.

Companion web site. The paper's companion web site (<http://worms.ucsf.edu/compare>) allows users to analyze their own microarray data sets using the tools in this paper, dynamically explore the paper's database

search results, identify significant GO categories associated with each search result, and browse the associated genes and measurements.

UNIVERSITY OF CALIFORNIA
LIBRARY
UNIVERSITY OF CALIFORNIA
LIBRARY
UNIVERSITY OF CALIFORNIA
LIBRARY
UNIVERSITY OF CALIFORNIA
LIBRARY
UNIVERSITY OF CALIFORNIA
LIBRARY
UNIVERSITY OF CALIFORNIA
LIBRARY

UNIVERSITY OF CALIFORNIA
LIBRARY



References

- Alter, O., Brown, P. O., and Botstein, D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A* 100, 3351-3356.
- Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297, 2270-2275.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29.
- Bluher, M., Kahn, B., and Kahn, C. (2003). Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science* 299, 572-574.
- Boldrick, J. C., Alizadeh, A. A., Diehn, M., Dudoit, S., Liu, C. L., Belcher, C. E., Botstein, D., Staudt, L. M., Brown, P. O., and Relman, D. A. (2002). Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc Natl Acad Sci U S A* 99, 972-977.
- Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G., *et al.* (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 31, 68-71.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* 282, 699-705.

Chung, C. H., Bernard, P. S., and Perou, C. M. (2002). Molecular portraits and the family tree of cancer. *Nat Genet* 32 *Suppl*, 533-540.

Cuadras, M. A., Feigelstock, D. A., An, S., and Greenberg, H. B. (2002). Gene expression pattern in Caco-2 cells following rotavirus infection. *J Virol* 76, 4467-4482.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-686.

Detweiler, C. S., Cunanan, D. B., and Falkow, S. (2001). Host microarray analysis reveals a role for the Salmonella response regulator phoP in human macrophage cell death. *Proc Natl Acad Sci U S A* 98, 5850-5855.

Dillin, A., Crawford, D. K., and Kenyon, C. (2002a). Timing requirements for insulin/IGF-1 signaling in *C. elegans*. *Science* 298, 830-834.

Dillin, A., Hsu, A. L., Arantes-Oliveira, N., Lehrer-Graiwer, J., Hsin, H., Fraser, A. G., Kamath, R. S., Ahringer, J., and Kenyon, C. (2002b). Rates of behavior and aging specified by mitochondrial function during development. *Science* 298, 2398-2401.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30, 207-210.

ci
RSIT
VERSI
LIB
m
n
ci
RSIT
VERSI
LIB
m
LIBOR
LIBOR
nci
R
NIVERS
NIVERS
18

LIBRARY



Finkel, T., and Holbrook, N. J. (2000). Oxidants, oxidative stress and the biology of ageing. *Nature* 408, 239-247.

Gaudet, J., and Mango, S. E. (2002). Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science* 295, 821-825.

Guillemin, K., Salama, N. R., Tompkins, L. S., and Falkow, S. (2002). Cag pathogenicity island-specific responses of gastric epithelial cells to *Helicobacter pylori* infection. *Proc Natl Acad Sci U S A* 99, 15136-15141.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17, 994-999.

Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G., and Brown, E. L. (2000). Genomic analysis of gene expression in *C. elegans*. *Science* 290, 809-812.

Hsin, H., and Kenyon, C. (1999). Signals from the reproductive system regulate the lifespan of *C. elegans*. *Nature* 399, 362-366.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., *et al.* (2000). Functional discovery via a compendium of expression profiles. *Cell* 102, 109-126.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., and Kim, S. K. (2001). Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* 98, 218-223.

Kayo, T., Allison, D. B., Weindruch, R., and Prolla, T. A. (2001). Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proc Natl Acad Sci U S A* 98, 5093-5098.

Kenyon, C., Chang, J., Gensch, E., Rudner, A., and Tabtiang, R. (1993). A *C. elegans* mutant that lives twice as long as wild type. *Nature* 366, 461-464.

Kimura, K. D., Tissenbaum, H. A., Liu, Y., and Ruvkun, G. (1997). *daf-2*, an insulin receptor-like gene that regulates longevity and diapause in *Caenorhabditis elegans*. *Science* 277, 942-946.

Lee, C. K., Klopp, R. G., Weindruch, R., and Prolla, T. A. (1999). Gene expression profile of aging and its retardation by caloric restriction. *Science* 285, 1390-1393.

Lewis, J. A., and Fleming, J. T. (1995). Basic Culture Methods. In *Methods in Cell Biology, Volume 48: Caenorhabditis elegans: Modern Biological Analysis of an Organism*, H. F. Epstein, and D. C. Shakes, eds. (San Diego, CA, Academic Press, Inc.), pp. 4-30.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nat Genet* 21, 20-24.

Lithgow, G. J., White, T. M., Melov, S., and Johnson, T. E. (1995). Thermotolerance and extended life-span conferred by single-gene mutations and induced by thermal stress. *Proc Natl Acad Sci U S A* 92, 7540-7544.

Lund, J., Tedesco, P., Duke, K., Wang, J., Kim, S. K., and Johnson, T. E. (2002). Transcriptional profile of aging in *C. elegans*. *Curr Biol* 12, 1566-1573.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al.* (2003). PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34, 267-273.

Murphy, C. T., McCarroll, S. A., Bargmann, C. I., Fraser, A., Kamath, R. S., Ahringer, J., Li, H., and Kenyon, C. (2003). Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* 424, 277-283.

Pletcher, S. D., Macdonald, S. J., Marguerie, R., Certa, U., Stearns, S. C., Goldstein, D. B., and Partridge, L. (2002). Genome-wide transcript profiles in aging and calorically restricted *Drosophila melanogaster*. *Curr Biol* 12, 712-723.

Raghavan, A., Ogilvie, R. L., Reilly, C., Abelson, M. L., Raghavan, S., Vasdevani, J., Krathwohl, M., and Bohjanen, P. R. (2002). Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res* 30, 5529-5538.

Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nat Genet* 33, 49-54.

Reinke, V., Smith, H. E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S. J., Davis, E. B., Scherer, S., Ward, S., and Kim, S. K. (2000). A global profile of germline gene expression in *C. elegans*. *Molecular Cell* 6, 605-616.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., *et al.* (2000). Comparative genomics of the eukaryotes. *Science* 287, 2204-2215.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34, 166-176.

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., Kaloper, M., Weng, S., Jin, H., Ball, C. A., *et al.* (2001). The Stanford Microarray Database. *Nucleic Acids Res* 29, 152-155.

Somani, S. M., Buckenmeyer, P., Dube, S. N., Mandalaywala, R. H., Verhulst, S. J., and Knowlton, R. G. (1992). Influence of age on caloric expenditure during exercise. *Int J Clin Pharmacol Ther Toxicol* 30, 1-6.

Stoeckert, C. J., Jr., Causton, H. C., and Ball, C. A. (2002). Microarray databases: standards and ontologies. *Nat Genet* 32 *Suppl*, 469-473.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249-255.

Teichmann, S. A., and Babu, M. M. (2002). Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol* 20, 407-410.

van Noort, V., Snel, B., and Huynen, M. A. (2003). Predicting gene function by conserved co-expression. *Trends Genet* 19, 238-242.

Wang, D. Y., Kumar, S., and Hedges, S. B. (1999). Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc R Soc Lond B Biol Sci* 266, 163-171.

Wang, W., Cherry, J. M., Botstein, D., and Li, H. (2002a). A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 99, 16893-16898.

Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., and Brown, P. O. (2002b). Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* 99, 5860-5865.

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O., and Botstein, D.

RY
IVERSIT
IVERSI
IBI
an
VERSIT
ERSI
an
VERS
VERS
an

VERSIT
ERSI
an
VERS
VERS
an



(2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* 13, 1977-2000.

Whitney, A. R., Diehn, M., Popper, S. J., Alizadeh, A. A., Boldrick, J. C., Relman, D. A., and Brown, P. O. (2003). Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci U S A* 100, 1896-1901.

Zou, S., Meadows, S., Sharp, L., Jan, L. Y., and Jan, Y. N. (2000). Genome-wide study of aging and oxidative stress response in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* 97, 13726-13731.

Table 1. Closest matches among all *C. elegans* microarray experiments to an expression profile of aging in *D. melanogaster*. (heads from 47-day-old vs. heads from 3-day-old male flies). The database includes gene expression profiles addressing larval development, germ line development, adult aging, environmental stress responses, neuronal signaling, and cell fate defects. The database was made from about 340 microarray experiments, of which 40 are profiles of aging.

<i>C. elegans</i> samples compared	Correlation to profile of aging in <i>Drosophila</i>	Ortholog pairs	P
2-weeks-old vs. 2.5 days old (Hill et al., 2000)	0.180	408	8.34×10^{-5}
Adult-age-144hr vs. Adult-age-0hr	0.148	2040	9.24×10^{-12}
Adult-age-52hr vs. Adult-age-0hr	0.145	2037	2.45×10^{-11}
Adult-age-40hr vs. Adult-age-0hr	0.139	1902	5.67×10^{-10}
Adult-age-72hr vs. Adult-age-0hr	0.134	2029	6.74×10^{-10}
Adult-age-16hr vs. Adult-age-0hr	0.122	1934	3.74×10^{-8}
Adult-age-96hr vs. Adult-age-0hr	0.119	1993	5.18×10^{-8}
Adult-age-8hr vs. Adult-age-0hr	0.115	1947	1.77×10^{-7}
Adult-age-28hr vs. Adult-age-0hr	0.092	1965	2.24×10^{-5}
Adult-age-52hr vs. Adult-age-8hr	0.091	2000	2.34×10^{-5}
Adult-age-52hr vs. Adult-age-16hr	0.089	2055	2.66×10^{-5}
Heat-stressed 2 hr. vs. Control	0.089	2129	1.97×10^{-5}
Adult-age-144hr vs. Adult-age-8hr	0.086	2010	5.66×10^{-5}
Adult-age-72hr vs. Adult-age-8hr	0.085	1998	7.13×10^{-5}
Heat-stressed 4 hr. vs. Control	0.083	2118	6.57×10^{-5}
.....			
320 profiles of varied biological phenomena	-0.06 to 0.06		
.....			
<i>daf-2</i> (<i>e1368</i>) mutants vs. Wild-type (N2) (Murphy et al., 2003)	-0.132	2156	3.67×10^{-10}

ucis
RY
IVERSIT
IVERSI
181
m
L
L
nci
R
IVERSIT
IVERSI
181
m
L
L
LIFORN
LIFORI
nci
R
NIVERS
NIVERS
181
m

181
m
L
L



Table 2. Cross-species searches of DNA microarray databases

<u>Query profile</u>	<u>Gene expression database searched</u>	<u>Matching profiles (best hits)</u>	<u>Correlation (R)</u>	<u>Ortholog pairs</u>	<u>P</u>
<i>D. melanogaster</i> adult aging: d18 vs. d3 (Pletcher et al., 2002)	<i>C. elegans</i>	Aging: 16hr vs. 0hr adult age Aging: 2weeks vs. 2.5 days old (Hill et al., 2000) Aging: 72hr vs. 0hr adult age	0.152 0.150 0.139	2400 422 2505	4.47×10^{-15} 1.04×10^{-4} 1.64×10^{-12}
<i>D. melanogaster</i> larval development: 96hrs. vs. 24 hrs. (Arbeitman et al., 2002)	<i>C. elegans</i>	Larval development: L2 vs. L1 (Jiang et al., 2001) Larval development: L3 vs. L1 (Jiang et al., 2001) Larval development: L4 vs. L1 (Jiang et al., 2001)	0.162 0.136 0.130	1334 1334 1334	1.27×10^{-9} 3.25×10^{-7} 8.83×10^{-7}
<i>D. melanogaster</i> embryonic development: 24hrs. vs. 11 hrs. (Arbeitman et al., 2002)	<i>C. elegans</i>	Embryonic/larval development: 12h vs. egg (Hill et al., 2000) Embryonic/larval development: L1 vs. egg (Jiang et al., 2001) Embryonic development: <i>skn-1</i> vs. <i>par-1</i> embryos (Gaudet and Mango, 2002)	0.306 0.214 0.176	624 1284 800	3.04×10^{-14} 2.00×10^{-12} 5.90×10^{-8}
<i>S. cerevisiae</i> sporulation: t=2hr vs. t=0hr (Chu et al., 1998)	<i>C. elegans</i>	Germ line: N2 vs. <i>glp-4</i> , young adults (Reinke et al., 2000) Germ line: N2 vs. <i>glp-4</i> , L3s (Reinke et al., 2000(i)) Germ line: N2 vs. <i>glp-4</i> , L2s (Reinke et al., 2000)	0.121 0.098 0.082	730 629 713	5.00×10^{-4} 6.00×10^{-3} 1.40×10^{-2}
<i>H. sapiens</i> mRNA decay: actinomycin D 45 min. vs. baseline (Raghavan et al., 2002)	<i>S. cerevisiae</i>	mRNA decay: <i>rpb1</i> vs. wild-type, 10 min (Wang et al., 2002b) mRNA decay: <i>rpb1</i> vs. wild-type, 5 min mRNA decay: <i>rpb1</i> vs. wild-type, 15 min	0.322 0.321 0.317	719 717 719	4.13×10^{-19} 5.32×10^{-19} 1.50×10^{-18}

UNIVERSITY OF
MICHIGAN
LIBRARY

UNIVERSITY OF
MICHIGAN
LIBRARY

8
ST
SI
R
n
C
L
ST
SI
R
n
C
L
ST
SI
R
n
C
L
ST
SI
R
n
C
L

1940

Figure 1. Comparative functional genomic analysis schematic.

(a) Phylogenetic analysis. Comparative sequence analysis is used to identify a complete set of candidate orthologs -- genes related by vertical descent from a gene (asterisk) present in the last common ancestor of *C. elegans* and *Drosophila*.

If an orthologous group has multiple genes from either species, the most-conserved orthologous gene pair is identified (for example, from group 3851, which contains the two *Drosophila* paralogs *Sep1* and *pnut*, *C. elegans* *unc-59* and *Drosophila* *pnut* were selected).

(b) Expression profiling. For each organism, DNA microarrays are used to measure the relative expression of each gene under two conditions.

(c) Phylogenetic integration of expression data. Measurements of log-fold-change in expression are systematically paired between orthologous genes from the two organisms. The correlation of the paired log-fold-change measurements is used to assess the similarity of the gene expression patterns in the two organisms. Hypothetical data is used here to illustrate the fact that even if most ortholog pairs (grey circles) lack any conserved regulation and contribute no correlation, a set of ortholog pairs (black circles) with partially conserved regulation can create a significant global correlation. For the data shown, in which conserved regulation contributes to expression of 25% of the ortholog pairs, the global Pearson correlation $r = 0.15$.

UNIVERSITY OF
MICHIGAN
LIBRARY

UNIVERSITY OF MICHIGAN LIBRARY



Figure 1

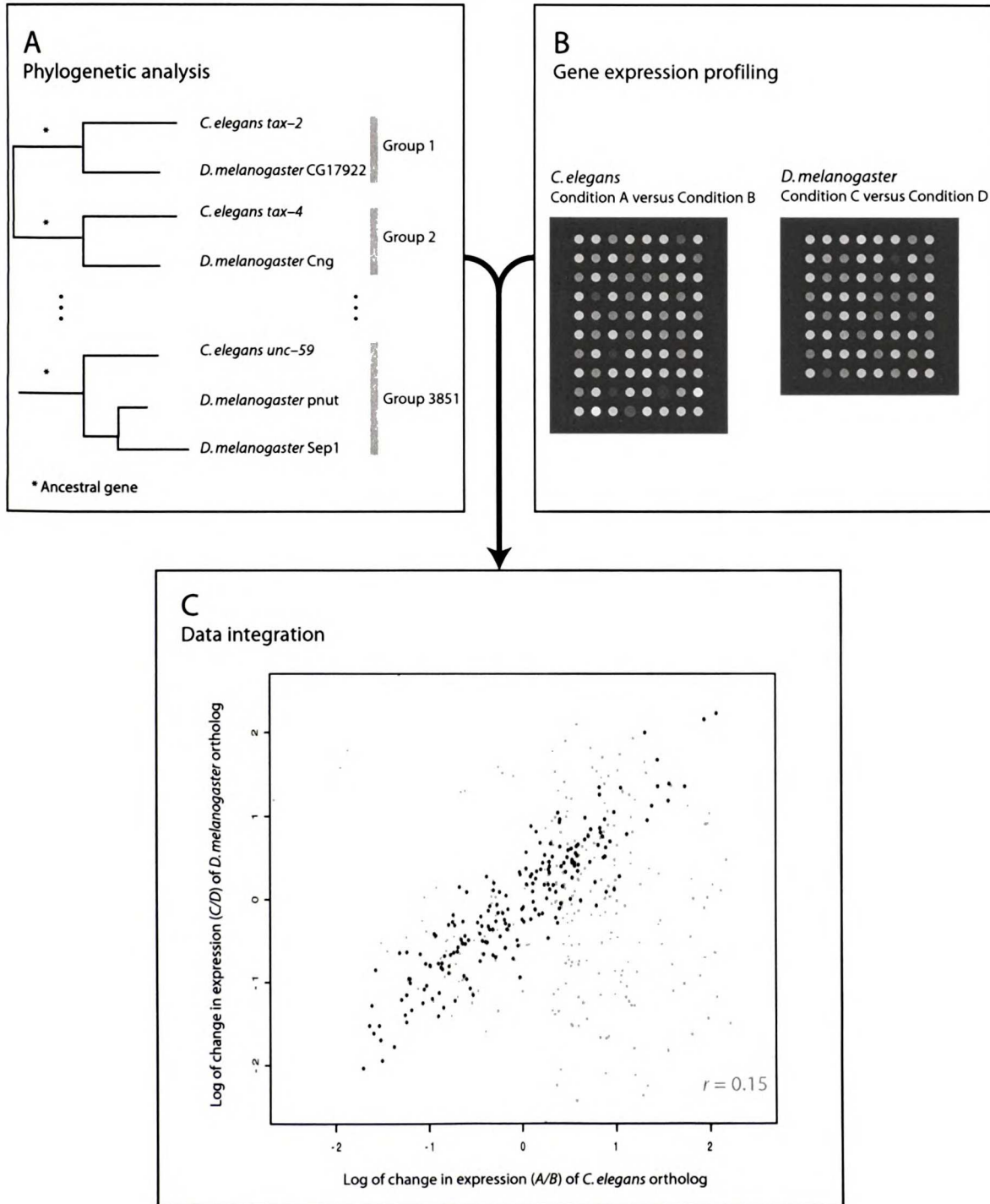


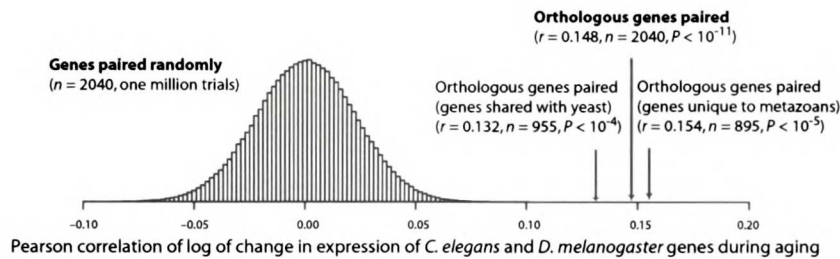
Figure 2. Correlated regulation of orthologous genes by aging in *C. elegans* and *Drosophila*.

(a) Correlated effect of aging on expression of orthologous genes in *C. elegans* and *D. melanogaster*. Microarray measurements of log-fold-change in expression with age were paired for orthologous genes from *C. elegans* and *D. melanogaster*; this Pearson correlation for orthologous gene pairs (long arrow) was compared against a distribution of one million Pearson correlations (histogram) each obtained by pairing *C. elegans* and *Drosophila* genes randomly.

(b) Shared transcriptional signature of aging in *D. melanogaster* heads and *C. elegans*. The methodology presented here (see Methods, Gene Ontology analysis) was used to identify highly conserved patterns within the gene expression data sets (fourteen large blocks in figure) that corresponded to Gene Ontology (GO) categories. For each GO category, the measured change in expression of each gene (small colored rectangle within block) in that GO category is represented. Each *D. melanogaster* gene is shown above its *C. elegans* ortholog. Red indicates induction by aging; green indicates repression by aging. All ortholog pairs from the significant GO categories are shown; some GO categories overlap, with some ortholog pairs belonging to more than one category. Statistical inferences have been made at the GO category level; most individual genes show small or statistically insignificant fold-changes, but the broad pattern of these changes is conserved and highly significant. **Supplementary Fig. 1** allows users to inspect the identities of the individual genes whose expression is represented in the figure.

Figure 2

A



B

Cellular components

Mitochondrial membrane (GO:0005740, $P < 10^{-4}$)



Mitochondrial inner membrane (GO:0005743, $P < 10^{-4}$)



Molecular functions

Carrier (GO:0005386, $P < 10^{-7}$)



Primary active transporter (GO:0015399, $P < 10^{-6}$)



Hydrolase, acting on acid anhydrides, - catalysing transmembrane movement of substances (GO:0016820, $P < 10^{-5}$)



P-P-bond-hydrolysis-driven transporter (GO:0015405, $P < 10^{-4}$)



Hydrolase, acting on carbon-nitrogen (but not peptide) bonds (GO:0016810, $P < 10^{-4}$)



Ion transporter (GO:0015075, $P < 10^{-3}$)



Cation transporter (GO:0008324, $P < 10^{-3}$)



ATP-binding cassette (ABC) transporter (GO:0004009, $P < 10^{-3}$)



Carbon-carbon lyase (GO:0016830, $P < 10^{-3}$)



Peptidase (GO:0008233, $P < 10^{-3}$)



DNA repair protein (GO:0003685, $P < 10^{-3}$)



Biological processes

Catabolism (GO:0009056, $P < 10^{-3}$)



Figure 3. Temporal distribution of conserved gene regulation across adulthood in *C. elegans* and *D. melanogaster*.

(a) Implementation of conserved gene expression changes over time. Genomic expression changes during six periods of *C. elegans* adulthood were compared to changes during four periods of *D. melanogaster* adulthood, by measuring the Pearson correlation of the log-fold-change in expression of orthologous genes.

Statistical significance of correlations: ** $p < 0.001$, * $p < 0.005$.

(b) Global correlations in genomic expression changes during aging in transcripts from whole *C. elegans* and *D. melanogaster* heads.

(c) Transcriptional repression of oxidative metabolism early in adulthood in *C. elegans* and *D. melanogaster*. Orthologous genes in the mitochondrial electron transport chain (GO:0005746) were identified. The expression of each gene, relative to its expression at the beginning of the time course, was obtained from the microarray data sets. The figure shows the median fold-change (connected points) and two standard errors around the geometric mean fold-change (bars) for this group of genes.

(d) Implementation of a stress response pattern early in adulthood in *C. elegans* and *D. melanogaster*. Profiles of gene regulation during successive periods of *C. elegans* and *D. melanogaster* adulthood were compared to profiles of *C. elegans* heat stress (light bars) and *D. melanogaster* oxidative stress (dark bars), by measuring the Pearson correlation of the log-fold-regulation of orthologous genes (for inter-species comparisons) or of the same genes (for same-species comparisons). Microarray experiments are described in Methods. Statistical significance of correlations: * $p < 0.001$.

RY
UNIVERSIT
UNIVERSI
181
n
UNIVERSIT
UNIVERSI
181
n
UNIVERSIT
UNIVERSI
181
n
UNIVERSIT
UNIVERSI
181
n

181
n
UNIVERSIT
UNIVERSI
181
n
UNIVERSIT
UNIVERSI
181
n

Figure 3

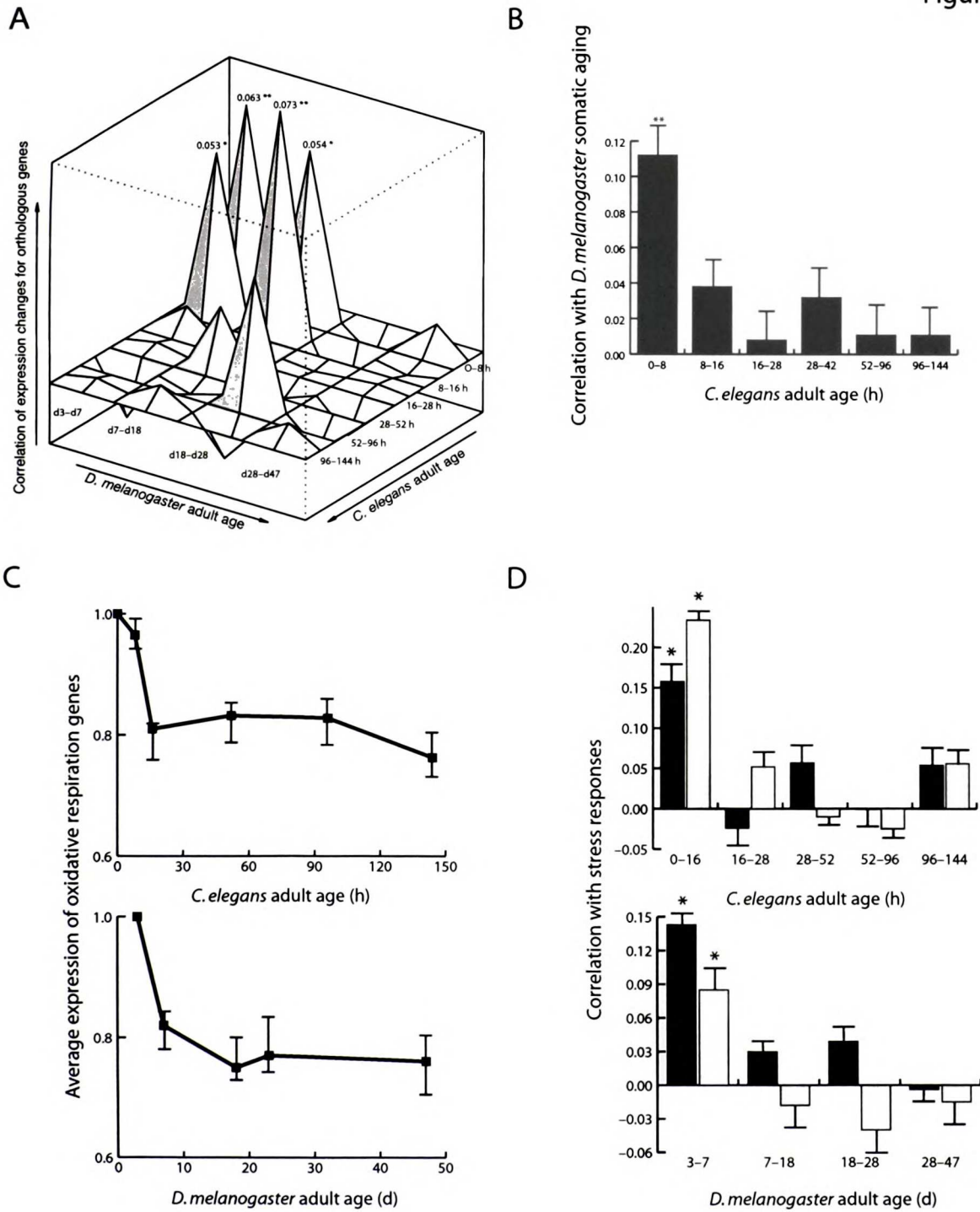


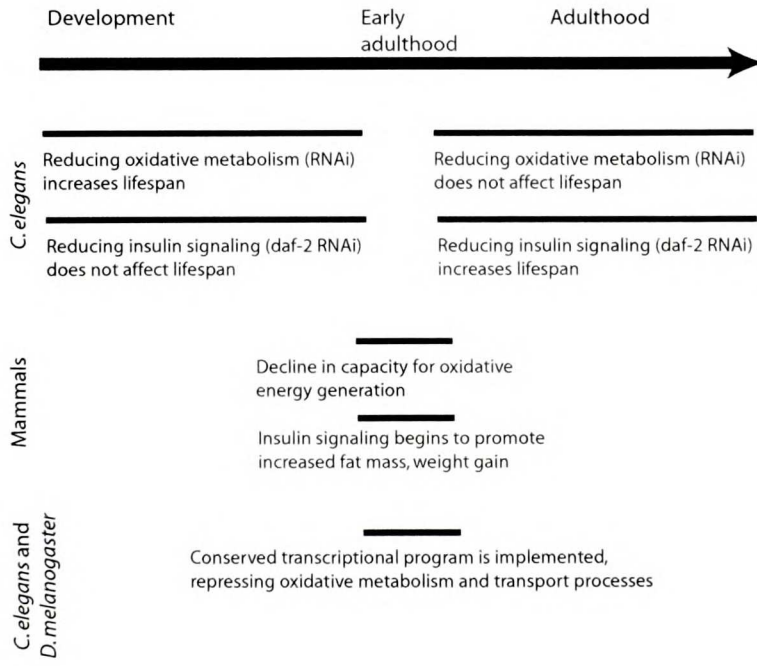
Figure 4. Aging, lifespan, and conserved early-adult physiological change in metazoa.

The insulin pathway begins to regulate lifespan on the first day of adulthood in *C. elegans* (Dillin et al., 2002a). In mice, insulin signaling in adipose tissue starts to cause weight gain early in adulthood (Blucher et al., 2003). In *C. elegans*, oxidative respiration appears to be lifespan-limiting until early adulthood, but not afterwards (Dillin et al., 2002b). Mammals begin to lose capacity for oxidative energy generation early in adulthood (Somani et al., 1992). *C. elegans* and *D. melanogaster* implement a conserved transcriptional program early in adulthood, one feature of which is the repression of oxidative metabolism genes.

WEST LIBRARY

CALIFORNIA
San
LIB
UNIVERSITY
UNIVERSITY
FOR
CALIFORNIA
San
LIB
UNIVERSITY
UNIVERSITY
LIBRARY
ncis
CALIFORNIA
CALIFORNIA
San
LIB
UNIVERSITY

Figure 4



Supplementary Figure 1

Transcriptional signature of aging conserved between *C. elegans* and *D. melanogaster* heads

Each large block in the figure below corresponds to a Gene Ontology category in which orthologous genes showed significantly conserved patterns of change in expression in these microarray experiments. Each small colored rectangle within each block represents measurement of change in expression for a specific gene. Orthologous genes are vertically paired within each block. The Gene Ontology categories can overlap, with some ortholog pairs belonging to more than one category. When this figure is viewed in a web browser, each of the small colored rectangles links to corresponding gene information in genome databases. Statistical inferences have been made at the GO category level; most individual genes show small or statistically insignificant fold-changes, but the broad pattern of these changes is conserved and highly significant.

The figure is the direct output of the paper's methodology for identifying the Gene Ontology categories associated with conserved patterns of gene regulation (see Methods, Gene Ontology analysis). The paper's companion web site, <http://worms.ucsf.edu/compare>, allows researchers to perform such analyses on any pair of microarray data sets from different organisms.

Experiment 1 *Drosophila melanogaster* Aging: heads_d47_vv_d3 (see Methods)

Experiment 2 *Caenorhabditis elegans* Aging: Adult144hr_vv_Adult0hr (see Methods)

2043 ortholog pairs, $r = 0.148$

CELLULAR COMPONENTS

Mitochondrial membrane (GO:0005740, $p < 10^{-4}$)



Mitochondrial inner membrane (GO:0005743, $p < 10^{-4}$)



MOLECULAR FUNCTIONS

Carrier (GO:0005386, $p < 10^{-7}$)



Primary active transporter (GO:0015399, $p < 10^{-6}$)



Hydrolase, acting on acid anhydrides, - catalysing transmembrane movement of substances (GO:0016820, $p < 10^{-5}$)



P-P-bond-hydrolysis-driven transporter (GO:0015405, $p < 10^{-4}$)



Hydrolase, acting on carbon-nitrogen (but not peptide) bonds (GO:0016810, $p < 10^{-4}$)



10

21

ERSIT

VERSI

BI

in

2

111

zch

R

IVERSIT

IVERSI

IB

in

2

ALIFORN

ALIFORN

anci

AR

NIVERSI

NIVERS

IB

in

Supplementary Figure 2

Transcriptional signatures of aging conserved between *C. elegans* and *D. melanogaster* in various gene expression experiments

The three analyses below apply the same analytical methodology (see Methods, Gene Ontology analysis) to different expression profiles of aging in *C. elegans* and *D. melanogaster*. The various experiments used different strains, culture conditions, and tissue types in each organism; nonetheless, the cross-species comparisons yield broadly overlapping results.

[C. elegans \(this paper, see Methods\) vs. D. melanogaster heads \(this paper, see Methods\)](#)
[C. elegans \(Hill et al., Science 290: 809-12, 2000\) vs. D. melanogaster heads \(this paper, see Methods\)](#)
[C. elegans \(this paper, see Methods\) vs. D. melanogaster \(Pletcher et al., Curr Biol 12: 712-23, 2002\)](#)

Each large block in the figure below corresponds to a Gene Ontology category in which orthologous genes showed significantly conserved patterns of change in expression in these microarray experiments. Each small colored rectangle within each block represents measurement of change in expression for a specific gene. Orthologous genes are vertically paired within each block. The Gene Ontology categories can overlap, with some ortholog pairs belonging to more than one category. When this figure is viewed in a web browser, each of the small colored rectangles links to corresponding gene information in genome databases. Statistical inferences have been made at the GO category level; most individual genes show small or statistically insignificant fold-changes, but the broad pattern of these changes is conserved and highly significant.

The figure is the direct output of the paper's methodology for identifying the Gene Ontology categories associated with conserved patterns of gene regulation (see Methods, Gene Ontology analysis). The paper's companion web site, <http://worms.ucsf.edu/compare>, allows researchers to perform such analyses on any pair of microarray data sets from different organisms.

Transcriptional signature conserved between aging in *C. elegans* (this paper, see Methods) and *D. melanogaster* heads (this paper, see Methods)

Experiment 1 *Drosophila melanogaster* Aging: heads_d47_vv_d3
Experiment 2 *Caenorhabditis elegans* Aging: Adult144hr_vv_Adult0hr
2043 ortholog pairs, $r = 0.148$

CELLULAR COMPONENTS

Mitochondrial membrane (GO:0005740, $p < 10^{-4}$)

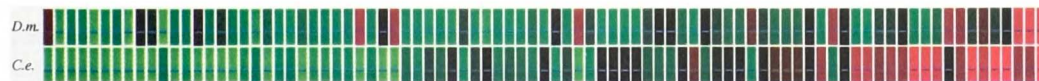


Mitochondrial inner membrane (GO:0005743, $p < 10^{-4}$)

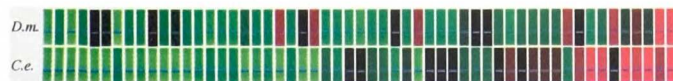


MOLECULAR FUNCTIONS

Carrier (GO:0005386, $p < 10^{-7}$)



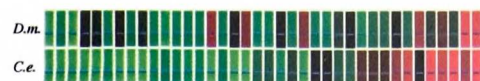
Primary active transporter (GO:0015399, $p < 10^{-6}$)



Hydrolase, acting on acid anhydrides, - catalysing transmembrane movement of substances (GO:0016820, $p < 10^{-5}$)



P-P-bond-hydrolysis-driven transporter (GO:0015405, $p < 10^{-4}$)



181
UNIVERSITY
OF
MICHIGAN
LIBRARY

Hydrolase, acting on carbon-nitrogen (but not peptide) bonds (GO:0016810, $p < 10^{-4}$)



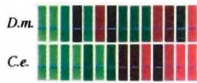
Ion transporter (GO:0015075, $p < 10^{-3}$)



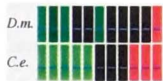
Cation transporter (GO:0008324, $p < 10^{-3}$)



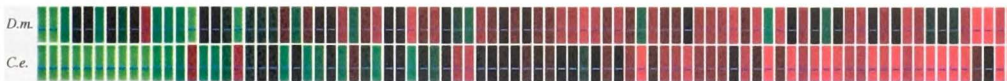
ATP-binding cassette (ABC) transporter (GO:0004009, $p < 10^{-3}$)



Carbon-carbon lyase (GO:0016830, $p < 10^{-3}$)



Peptidase (GO:0008233, $p < 10^{-3}$)



DNA repair protein (GO:0003685, $p < 10^{-3}$)



BIOLOGICAL PROCESSES

Catabolism (GO:0009056, $p < 10^{-3}$)



2x 2x

Repressed Induced
Key

Transcriptional signature conserved between aging in *C. elegans* (Hill et al., *Science* 290: 809-12, 2000) and *D. melanogaster* heads (this paper, see Methods)

Mitochondrial and transporter categories again show significantly conserved patterns in this comparison. Fewer overall categories show significant results, reflecting the fact that many *C. elegans* genes were not present in the Hill et al. data set, which reduced the number of ortholog pairs in the analysis.

Experiment 1 *Drosophila melanogaster* Aging: heads_d47_vv_d3

Experiment 2 *Caenorhabditis elegans* Aging: 2week_vv_60h

408 ortholog pairs, $r = 0.180$

110

110

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

UNIVERSITY

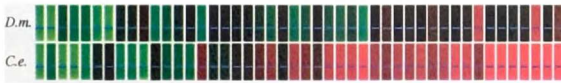
UNIVERSITY



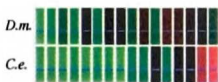
Mitochondrial electron transport chain complex (GO:0005746, $p < 10^{-6}$)



Cytoskeleton (GO:0005856, $p < 10^{-4}$)



Extracellular (GO:0005576, $p < 10^{-4}$)



Mitochondrial inner membrane (GO:0005743, $p < 10^{-3}$)

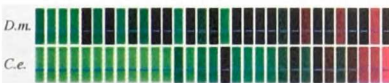


MOLECULAR FUNCTIONS

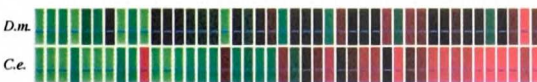
Peptidase (GO:0008233, $p < 10^{-6}$)



Metallopeptidase (GO:0008237, $p < 10^{-6}$)



Calcium binding (GO:0005509, $p < 10^{-6}$)



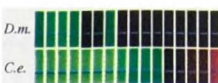
Transmembrane receptor (GO:0004888, $p < 10^{-6}$)



Receptor (GO:0004872, $p < 10^{-5}$)

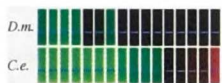


Metalloexopeptidase (GO:0008235, $p < 10^{-5}$)



1871

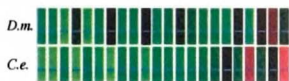
Exopeptidase (GO:0008238, $p < 10^{-5}$)



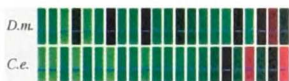
Cysteine-type peptidase (GO:0008234, $p < 10^{-4}$)



Hydrogen ion transporter (GO:0015078, $p < 10^{-4}$)



Monovalent inorganic cation transporter (GO:0015077, $p < 10^{-4}$)



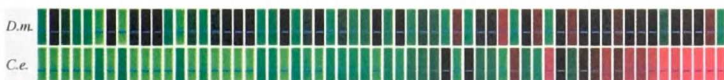
Cysteine-type endopeptidase (GO:0004197, $p < 10^{-4}$)



Cation transporter (GO:0008324, $p < 10^{-3}$)



Primary active transporter (GO:0015399, $p < 10^{-3}$)

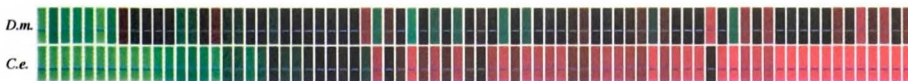


Ion transporter (GO:0015075, $p < 10^{-3}$)

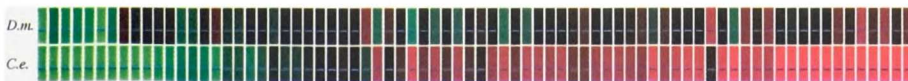


BIOLOGICAL PROCESSES

Protein degradation (GO:0030163, $p < 10^{-7}$)



Proteolysis and peptidolysis (GO:0006508, $p < 10^{-7}$)



Macromolecule catabolism (GO:0009057, $p < 10^{-7}$)

Supplementary Figure 3

Transcriptional signature of larval development conserved between *C. elegans* and *D. melanogaster*

Each large block in the figure below corresponds to a Gene Ontology category in which orthologous genes showed significantly conserved patterns of change in expression in these microarray experiments. Each small colored rectangle within each block represents measurement of change in expression for a specific gene. Orthologous genes are vertically paired within each block. The Gene Ontology categories can overlap, with some ortholog pairs belonging to more than one category. When this figure is viewed in a web browser, each of the small colored rectangles links to corresponding gene information in genome databases. Statistical inferences have been made at the GO category level; most individual genes show small or statistically insignificant fold-changes, but the broad pattern of these changes is conserved and highly significant.

The figure is the direct output of the paper's methodology for identifying the Gene Ontology categories associated with conserved patterns of gene regulation (see Methods, Gene Ontology analysis). The paper's companion web site, <http://worms.ucsf.edu/compare>, allows researchers to perform such analyses on any pair of microarray data sets from different organisms.

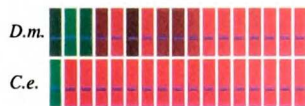
Experiment 1 *Drosophila melanogaster* Larval development 96hr_vv_24hr

Experiment 2 *Caenorhabditis elegans* Larval development: L2_vv_L1

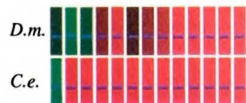
1334 ortholog pairs, $r = 0.162$

CELLULAR COMPONENTS

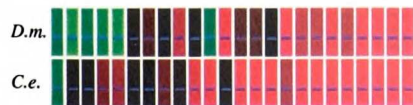
26S proteasome (GO:0005837, $p < 10^{-10}$)



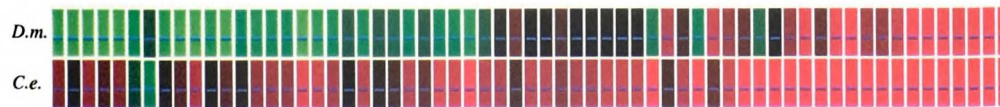
19S proteasome regulatory particle (GO:0005838, $p < 10^{-10}$)



Coated vesicle (GO:0030135, $p < 10^{-5}$)



Cytosol (GO:0005829, $p < 10^{-4}$)



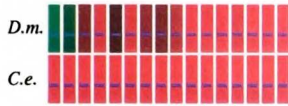
Cytoplasmic vesicle (GO:0016023, $p < 10^{-4}$)



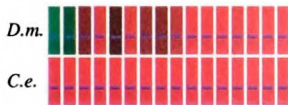


MOLECULAR FUNCTIONS

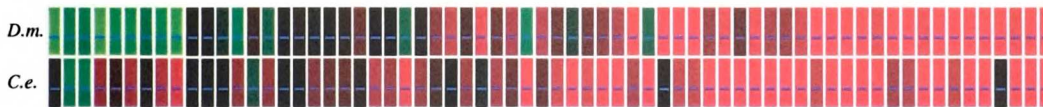
Threonine endopeptidase (GO:0004298, $p < 10^{-10}$)



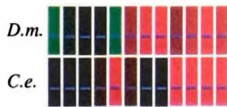
Proteasome endopeptidase (GO:0004299, $p < 10^{-10}$)



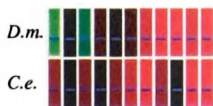
Peptidase (GO:0008233, $p < 10^{-10}$)



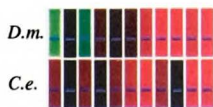
Isomerase (GO:0016853, $p < 10^{-5}$)



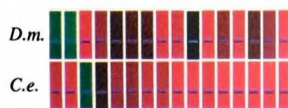
Metalloexopeptidase (GO:0008235, $p < 10^{-4}$)



Exopeptidase (GO:0008238, $p < 10^{-4}$)

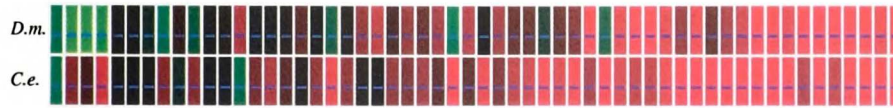


Hydrolase, acting on acid anhydrides, - involved in cellular and subcellular movement (GO:0016821, $p < 10^{-3}$)

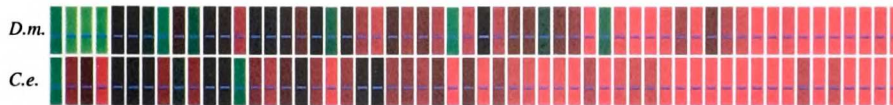


BIOLOGICAL PROCESSES

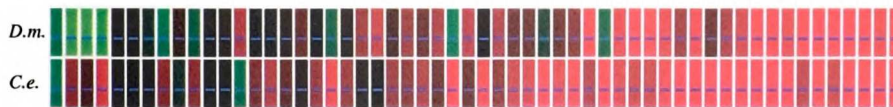
Macromolecule catabolism (GO:0009057, $p < 10^{-9}$)



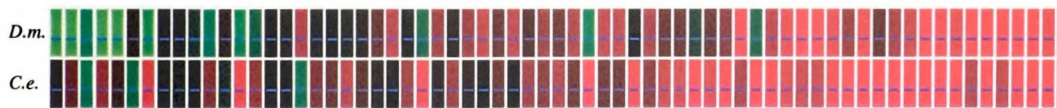
Protein degradation (GO:0030163, $p < 10^{-9}$)



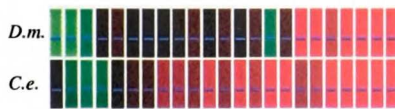
Proteolysis and peptidolysis (GO:0006508, $p < 10^{-9}$)



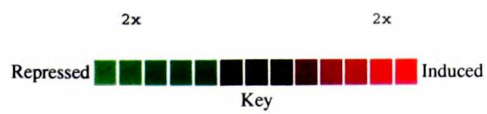
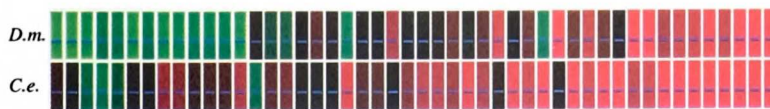
Catabolism (GO:0009056, $p < 10^{-8}$)



Intracellular protein transport (GO:0006886, $p < 10^{-4}$)



Protein transport (GO:0015031, $p < 10^{-3}$)



Supplementary Figure 4

Transcriptional signature of embryonic development conserved between *C. elegans* and *D. melanogaster*

Each large block in the figure below corresponds to a Gene Ontology category in which orthologous genes showed significantly conserved patterns of change in expression in these microarray experiments. Each small colored rectangle within each block represents measurement of change in expression for a specific gene. Orthologous genes are vertically paired within each block. The Gene Ontology categories can overlap, with some ortholog pairs belonging to more than one category. When this figure is viewed in a web browser, each of the small colored rectangles links to corresponding gene information in genome databases. Statistical inferences have been made at the GO category level; most individual genes show small or statistically insignificant fold-changes, but the broad pattern of these changes is conserved and highly significant.

The figure is the direct output of the paper's methodology for identifying the Gene Ontology categories associated with conserved patterns of gene regulation (see Methods, Gene Ontology analysis). The paper's companion web site, <http://worms.ucsf.edu/compare>, allows researchers to perform such analyses on any pair of microarray data sets from different organisms.

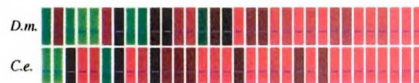
Experiment 1 *Drosophila melanogaster* [Embryonic development 24hr vv 11hr](#)

Experiment 2 *Caenorhabditis elegans* [Embryonic/larval development: L3 vv eeg](#)

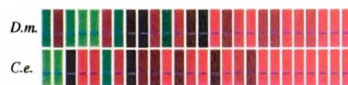
1284 ortholog pairs, $r = 0.218$

CELLULAR COMPONENTS

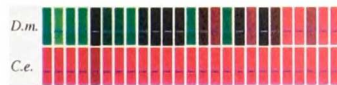
Mitochondrial membrane (GO:0005740, $p < 10^{-4}$)



Mitochondrial inner membrane (GO:0005743, $p < 10^{-4}$)



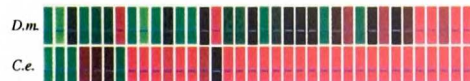
Cytosolic ribosome (sensu Eukarya) (GO:0005830, $p < 10^{-4}$)



Nucleoplasm (GO:0005654, $p < 10^{-4}$)



Ribosome (GO:0005840, $p < 10^{-3}$)



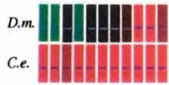
Mitochondrial matrix (GO:0005759, $p < 10^{-3}$)



Cytosolic small ribosomal subunit (sensu Eukarya) (GO:0005843, $p < 10^{-3}$)

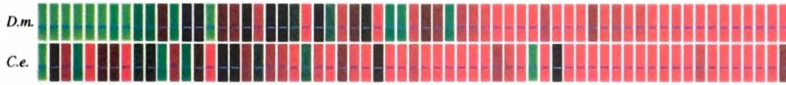


Eukaryotic 48S initiation complex (GO:0016283, $p < 10^{-3}$)

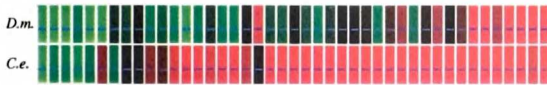


MOLECULAR FUNCTIONS

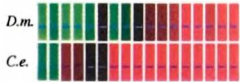
Oxidoreductase (GO:0016491, $p < 10^{-10}$)



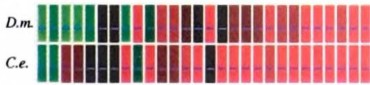
Structural molecule (GO:0005198, $p < 10^{-10}$)



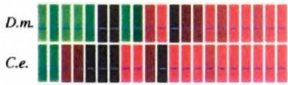
Hydrolase, acting on acid anhydrides, - catalysing transmembrane movement of substances (GO:0016820, $p < 10^{-6}$)



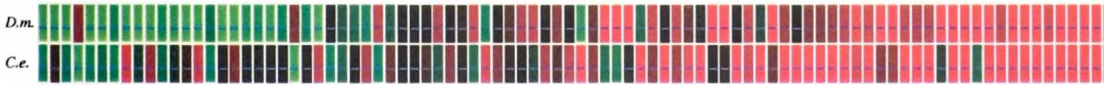
Primary active transporter (GO:0015399, $p < 10^{-6}$)



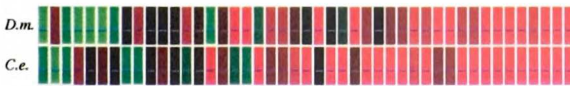
P-P-bond-hydrolysis-driven transporter (GO:0015405, $p < 10^{-6}$)



Transporter (GO:0005215, $p < 10^{-6}$)



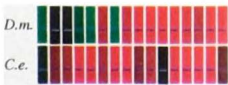
Carrier (GO:0005386, $p < 10^{-5}$)



Electron transporter (GO:0005489, $p < 10^{-5}$)

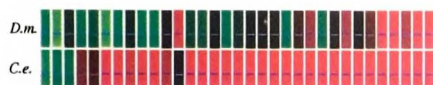


Oxidoreductase, acting on the CH-OH group of donors, NAD or NADP as acceptor (GO:0016616, $p < 10^{-4}$)

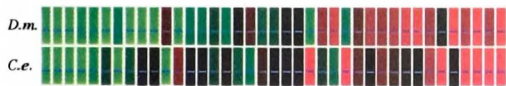


Structural constituent of ribosome (GO:0003735, $p < 10^{-4}$)

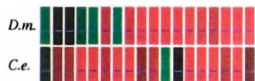




Cytoskeletal protein binding (GO:0008092, $p < 10^{-3}$)

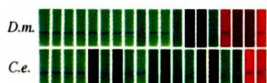


Oxidoreductase, CH-OH group of donors (GO:0016614, $p < 10^{-3}$)

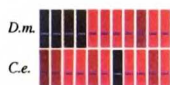


BIOLOGICAL PROCESSES

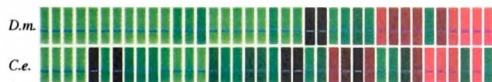
Mitotic cell cycle (GO:0000278, $p < 10^{-9}$)



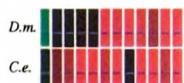
Catabolic carbohydrate metabolism (GO:0006095, $p < 10^{-6}$)



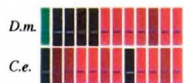
Cell cycle (GO:0007049, $p < 10^{-6}$)



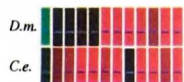
Energy pathways (GO:0006091, $p < 10^{-6}$)



Energy derivation by oxidation of organic compounds (GO:0015980, $p < 10^{-6}$)



Main pathways of carbohydrate metabolism (GO:0006092, $p < 10^{-6}$)



Amino acid and derivative metabolism (GO:0006519, $p < 10^{-5}$)



Locomotory behavior (GO:0007626, $p < 10^{-4}$)

LIBRARY
UNIVERSITY OF
ALBANY
STATE UNIVERSITY OF
NEW YORK
LIBRARY
UNIVERSITY OF
ALBANY
STATE UNIVERSITY OF
NEW YORK
LIBRARY
UNIVERSITY OF
ALBANY
STATE UNIVERSITY OF
NEW YORK

Supplementary Figure 5

Transcriptional signature of gametogenesis conserved between *S. cerevisiae* sporulation and *C. elegans* germ line formation

Each large block in the figure below corresponds to a Gene Ontology category in which orthologous genes showed significantly conserved patterns of change in expression in these microarray experiments. Each small colored rectangle within each block represents measurement of change in expression for a specific gene. Orthologous genes are vertically paired within each block. The Gene Ontology categories can overlap, with some ortholog pairs belonging to more than one category. When this figure is viewed in a web browser, each of the small colored rectangles links to corresponding gene information in genome databases. Statistical inferences have been made at the GO category level; most individual genes show small or statistically insignificant fold-changes, but the broad pattern of these changes is conserved and highly significant.

The figure is the direct output of the paper's methodology for identifying the Gene Ontology categories associated with conserved patterns of gene regulation (see Methods, Gene Ontology analysis). The paper's companion web site, <http://worms.ucsf.edu/compare>, allows researchers to perform such analyses on any pair of microarray data sets from different organisms.

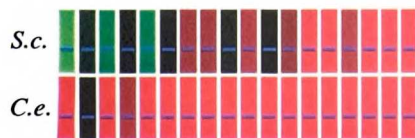
Experiment 1 *Saccharomyces cerevisiae* [Sporulation_t2hrs_vv_Sporulation_t0](#)

Experiment 2 *Caenorhabditis elegans* [Germ line: N2_vv_glp4_T3_avg](#)

720 ortholog pairs, $r = 0.121$

CELLULAR COMPONENTS

Replication fork (GO:0005657, $p < 10^{-8}$)

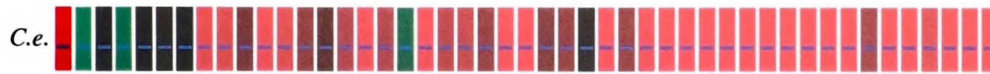


Nucleoplasm (GO:0005654, $p < 10^{-3}$)



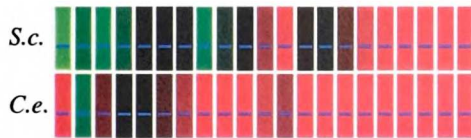
UNIVERSITY OF MICHIGAN LIBRARY

UNIVERSITY OF MICHIGAN LIBRARY

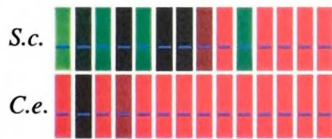


BIOLOGICAL PROCESSES

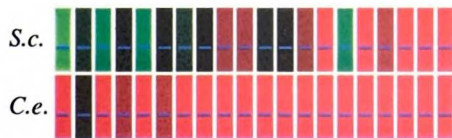
DNA repair (GO:0006281, $p < 10^{-6}$)



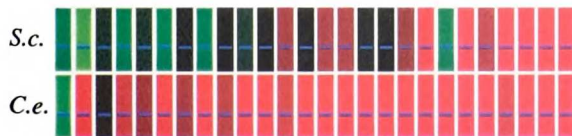
DNA strand elongation (GO:0006271, $p < 10^{-6}$)



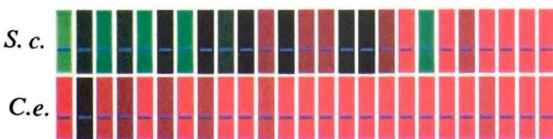
DNA dependent DNA replication (GO:0006261, $p < 10^{-5}$)



S phase of mitotic cell cycle (GO:0000084, $p < 10^{-5}$)



DNA replication (GO:0006260, $p < 10^{-5}$)



DNA replication and chromosome cycle (GO:0000067, $p < 10^{-4}$)

LIBRARY
UNIVERSITY OF
MICHIGAN
ANN ARBOR
MICHIGAN
UNIVERSITY OF
MICHIGAN
ANN ARBOR
MICHIGAN
UNIVERSITY OF
MICHIGAN
ANN ARBOR
MICHIGAN

UNIVERSITY OF MICHIGAN
ANN ARBOR
MICHIGAN

18
RY
IVERSIT
IVERSIT
81
w
IV
ic
RY
IVERSIT
IVERSIT
81
w
ALFOR
ALFOR
nci
R
IVERSIT
IVERS
18
w

18
RY
IVERSIT
IVERSIT
81
w
ALFOR
ALFOR
nci
R
IVERSIT
IVERS
18
w

LIBRARY

UNIVERSITY OF CALIFORNIA
San Jose
LIBRARY

CHAPTER 2

COMING OF AGE IN *METAZOA*: WHAT YOUNG ADULTHOOD CAN TELL US ABOUT AGING, LIFESPAN, METABOLISM, AND DISEASE

Abstract

Worms, flies, and humans implement a conserved adult-onset gene expression program. This program, which we call MYA (Metazoan Young Adult transcriptional program) includes the reduced expression of mitochondrial oxidative metabolism and cellular transport processes. MYA is implemented in somatic tissues, including human skeletal muscle, *Drosophila* heads, and rat hippocampal neurons. Understanding MYA may help to reconcile disparate findings in aging and lifespan research. MYA may also affect the course and presentation of adult-onset human illnesses, such as type II diabetes and hypertension. Understanding the regulation and consequences of conserved adult-onset changes in gene expression presents an exciting research direction.

Introduction

The study of early adulthood -- the period after an animal reaches reproductive maturity and before it reaches middle age -- has dwelt neglected at the intersection between two active fields: developmental biology and aging research. Because early adulthood takes place after changes in body size and shape, it has not been a central concern in developmental biology; because early adulthood precedes the appearance of aging phenotypes, early adulthood has not been a central concern in aging research. The observation of a widely

conserved, early-adult-onset gene expression program should enhance research interest in this transitional period in the metazoan life cycle.

Adulthood in animals is generally considered a period of developmental stasis characterized by reproduction, body plan stability, and aging. Actions by somatic cells in adult animals are often assumed not to be developmentally programmed, but rather to be responses to acute perturbations and cumulative degradation.

We recently found that the fruit fly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans* – two metazoan species separated by almost a billion years of evolution – execute a shared, developmentally-timed gene expression program early in adulthood (McCarroll et al., 2004). This shared program is broad, encompassing dozens to hundreds of genes, and it is detected in somatic tissue from *Drosophila* heads. Moreover, most of these conserved changes in gene expression are implemented early in adulthood – in *Drosophila*, between adult d3 and adult d7 in animals that will live for weeks longer, and in *C. elegans*, on the first day of adulthood in animals that will live for another 10-16 days. The early-adult implementation of this program suggests it is developmentally timed, rather than a graded response to cumulative damage or degeneration. This timing precedes the onset of aging phenotypes and corresponds approximately to periods of peak reproductive activity, though it is also observed in sterile animals (*C. elegans*) and in somatic tissue (*D. melanogaster*) and is therefore unlikely to correspond to reproduction itself (McCarroll et al., 2004). Although this result is surprising, its interpretation requires answers to several open questions. First, because the transcriptional programs were measured from mixed-tissue preparations, it is difficult to assign the program to

specific somatic tissues. Second, because of the relatively short life cycles of *C. elegans* and *D. melanogaster*, it is difficult to definitively exclude the possibility that it represents a delayed implementation of changes associated with the termination of development. Third, although worms and flies represent an ancient divergence of the metazoan family, it cannot be assumed that these processes are conserved in mammals.

Here we find, using published data sets, that a similar adult-onset gene expression program is implemented in human skeletal muscle and rat hippocampal neurons. The mammalian gene expression program is implemented during periods of adulthood that are well separated from canonical developmental processes, in specific, well-defined somatic tissues. We discuss the implications of this finding for thinking about young adulthood and its relevance to aging, lifespan, metabolism, and disease.

A conserved metazoan young adult (MYA) transcriptional program

To assess whether young-adult patterns of transcriptional change shared by *C. elegans* and *D. melanogaster* were also conserved in humans, we used published gene expression profiles of aging in human skeletal muscle. Gene expression in *vastus lateralis* (quadriceps) muscle biopsies from seven young adult (20-29 years of age) and eight aged adult (65-71 years of age) women has been profiled using Affymetrix microarrays (Welle et al., 2004). Comparing the young-adult and aged-adult gene expression patterns allows the measurement of patterns of change in muscle gene expression during human adulthood. For each gene, a measurement of aged-adult or young-adult expression was obtained by taking the median normalized expression level across all of the aged-adult or young-adult samples. To measure the adult-progression change in expression of

RY
UNIVERSIT
UNIVERSI
LIB
un
L
UNIVERSIT
UNIVERSI
LIB
un
L
ALFORN
ALFOR
uncl
AR
UNIVERSI
UNIVERS
LIB
un

LIBRARY
UNIVERSITY OF CALIFORNIA
LIBRARY
UNIVERSITY OF CALIFORNIA
LIBRARY
UNIVERSITY OF CALIFORNIA



each gene, these composite young-adult and aged-adult profiles were compared. This profile of adult-progression change in gene expression was then compared to similar profiles obtained earlier in *D. melanogaster* ((McCarroll et al., 2004; Pletcher et al., 2002) and *C. elegans* (Hill et al., 2000; McCarroll et al., 2004), by measuring the Pearson correlation of the log-changes in expression of orthologous genes.

Patterns of adult-onset change in gene expression in humans were significantly correlated to patterns of adult-onset change in *C. elegans* and *D. melanogaster* (Table 1). These correlations remained highly significant when subjected to two additional tests of significance. In the first test, the pairing of orthologous genes in the two organisms was randomly permuted, and the pairing-permuted correlations for the data sets were measured. Across one million comparisons of such pairing-permuted data sets, none of the pairing-permuted data sets yielded a correlation as great as that yielded when orthologous genes were paired. This allowed a high confidence level ($p < 10^{-6}$) to be assigned to the assertion orthologous genes showed significantly correlated patterns of regulation. A second significance test was used to assess whether the correlations can be significantly associated with the progression of adulthood itself, and not explained by a combination of chance and the potentially conserved regulon structure of the human and invertebrate genomes. In the second test (which requires many distinct, independently-profiled samples and was therefore applied to the data from Welle et al., 2004 and Pletcher et al, 2002), we permuted the assignment of clinical labels (“young adult”, “aged adult”) to the underlying biological samples and then repeated the analysis. Across all 200 thousand possible permuted assignments of the “young adult” and “aged adult”

labels to the biological samples, none yielded cross-species correlations as great as the actual assignment of samples; this allowed a high confidence level ($p < 10^{-5}$) to be assigned to the assertion that the correlated patterns are specifically associated with the progression of adulthood in humans and invertebrates.

Adult-onset gene expression programs in *C. elegans* and *D. melanogaster* have been shown to have in common the reduced expression of mitochondrial energy metabolism genes, and the reduced expression of transporter genes (McCarroll et al., 2004). Because worms and flies are such diverged metazoan species, we might expect all metazoans, including humans, to share these specific features. Alternatively, humans might share distinct patterns of gene expression change with arthropods and with nematodes. To identify functional categories of genes associated with conserved patterns of change in gene expression, we systematically searched hundreds of Gene Ontology categories for categories in which orthologous genes showed especially highly correlated adult-onset changes in expression, using the approach we developed earlier (McCarroll et al., 2004).

Similar to the comparison of *C. elegans* and *D. melanogaster*, conserved patterns of adult-onset change in gene expression in *H. sapiens* and *D. melanogaster* included reduced expression of mitochondrial oxidative phosphorylation and cellular transport processes (Figure 1). When patterns of adult-onset gene expression change from *H. sapiens* and *C. elegans* were compared, these same categories were identified; in fact the major patterns of change from all three organisms are readily aligned using the gene-ortholog relationships (Figure 2).

nce

RY

nci

nci

AR

UNIVERS

UNIVER

18

2

The major shared changes in adult-onset gene expression in *D. melanogaster* and *C. elegans* have been shown to be implemented in early adulthood, before middle age (McCarroll et al., 2004). Although the two-time-point design of the *H. sapiens* muscle experiment, with a single young-adult and a single aged-adult set of samples, does not allow us to determine whether the conserved changes are implemented before middle age, a recent publication of gene expression profiles from rat hippocampal neurons of young adult (4 months), middle-aged (14 months), and aged adult (24 months) animals (Blalock et al., 2003) does allow us to do so. We calculated from these data sets the distribution of gene expression changes for genes in the “oxidative phosphorylation” Gene Ontology category. As in *C. elegans* and *D. melanogaster*, reduction in expression of oxidative metabolism genes was mostly implemented before middle age, suggesting that the early-adult timing of these transcriptional changes is also conserved in mammals (Figure 3).

This pattern of adult-onset gene expression change is perhaps the most consistent and reproducible feature of the diverse genomic studies of aging which have been published (Jiang et al., 2001; Kayo et al., 2001; Lee et al., 1999; Lee et al., 2000; McCarroll et al., 2004; Pletcher et al., 2002; Weindruch et al., 2001). It is notably absent from those data sets in which researchers have excluded early adulthood in order to focus on changes that are implemented later (Lund et al., 2002). The absence of MYA from these data sets is a further, indirect confirmation of its early-adulthood timing.

MYA and aging

The early-adulthood timing of MYA suggests that MYA may be more a product of developmental biology than of aging. In fact, though MYA was

discovered in a comparison of gene expression patterns for aging, MYA might accurately be said to be less about aging than about coming of age. What potentially connects MYA to aging and lifespan research are two findings that central relationships in aging and lifespan are modulated at precisely this time.

Mitochondrial oxidative metabolism is lifespan-limiting in *C. elegans*, *D. melanogaster*, and *S. cerevisiae*. This evidence comes from the lifespan-enhancing effects of all of the following: reduction-of-function mutations in genes that encode mitochondrial proteins (Felkai et al., 1999; Feng et al., 2001); experiments in which the expression levels of mitochondrial components are targeted via RNA interference (Dillin et al., 2002; Lee et al., 2003); and mild doses of the mitochondrial toxin Actinomycin D (Dillin et al., 2002). The timing of action of this effect has recently been established using RNAi in *C. elegans*, with a surprising result: while reducing juvenile oxidative metabolism increases subsequent adult lifespan, reducing adult oxidative metabolism does not (Dillin et al., 2002). Researchers have proposed complex explanations for this finding, such as a “molecular memory” of juvenile metabolism that is later used to regulate adult aging. MYA presents a simple alternative explanation: oxidative metabolism may cease to be lifespan-limiting in adults precisely because adults use oxidative metabolism less aggressively than juveniles do. In fact, oxidative metabolism stops being lifespan-limiting after the first day of *C. elegans* adulthood, the day on which MYA’s endogenous reduction in oxidative metabolism genes is implemented. The higher juvenile levels of oxidative metabolism may exhaust the capacity of cells to detoxify reactive oxygen species.

Insulin signaling via the DAF-2/IGFR pathway reduces lifespan in *C. elegans* and *D. melanogaster* (Clancy et al., 2001; Kenyon et al., 1993; Kimura et al.,

1997; Tatar et al., 2001). This relationship, too, is transformed at the beginning of *C. elegans* adulthood: RNAi against *daf-2* in juveniles does not increase lifespan, while RNAi against *daf-2* in adults increases it significantly. The observation that the physiological consequences of insulin signaling change early in adulthood is also supported by experiments in mice. Mice with an adipose-tissue-specific knockout of the IGF-1 receptor show no apparent phenotype at eight weeks age (one week after the completion of puberty, but begin to show differences in adiposity from wild-type mice that are pronounced by twelve weeks age, and ultimately live much longer than wild-type animals (Bluher et al., 2003; Bluher et al., 2002). The contents of MYA are strongly correlated with the pattern of change in gene expression elicited by *daf-2* RNAi and seen in *daf-2* mutants, suggesting that this pathway may be modulated during MYA. In fact, a set of lifespan-promoting targets of *daf-16* was identified by comparing the expression of *daf-2* and control animals during early adulthood (Murphy et al., 2003).

How would the arrival of animals at reproductive maturity be communicated to somatic cells throughout the animal, allowing the implementation of MYA in somatic tissues? Signals from proliferating germ cells in the developing reproductive system limit lifespan in *C. elegans* and *D. melanogaster* and regulate the nuclear localization of the transcription factor DAF-16 in somatic cells (Arantes-Oliveira et al., 2002; Hsin and Kenyon, 1999; Lin et al., 2001). Such a signal might repress the implementation of MYA, allowing it to be implemented only after the reproductive system is mature.

Such a model, if confirmed experimentally, would be consistent with the “antagonistic pleiotropy” theory of aging, which invokes tradeoffs between the requirements of rapid reproductive maturation and subsequent adult lifespan

(Williams, 1957). According to this theory, evolution will tend to favor rapid development over subsequent adult health, so when tradeoffs exist between the two, evolution will tend to favor the faster-developing, shorter-lived solution (Williams, 1957). A potential corollary of the “antagonistic pleiotropy” model might be that once animals reach reproductive maturity, those tradeoffs can be renegotiated, potentially allowing the pursuit of a biological program more consistent with adult health, prolonged reproduction, and (in some species) prolonged caring for descendants. High levels of oxidative metabolism during development clearly speed the development of animals to reproductive maturity (the lifespan-enhancing effects of reductions in mitochondrial function all come at the cost of delayed development); but once animals have reached reproductive maturity, they appear to reduce oxidative metabolism to a level more consistent with adult longevity, at least in the sense that further experimental reductions do not further increase lifespan. In such a model, MYA might be seen as lifespan-enhancing, though its effect on aging may be complex – MYA might contribute to some specific physiological declines that are associated with aging, while preventing other declines by protecting tissues from oxidative and other damage.

Physiological significance of MYA

What are the physiological consequences of adult-onset changes in gene expression? Human variation in the expression levels of oxidative metabolism genes is highly correlated with total-body aerobic capacity (Mootha et al., 2003). Furthermore, OXPHOS gene expression reductions of the magnitude seen in MYA (20-30%) are similar to the difference observed between type II diabetics and healthy controls, suggesting that such changes are physiologically significant (Mootha et al., 2003). We speculate that MYA is associated with many well-

known adult-onset phenotypes, such as the increased adult-onset tendency to weight gain; the reduced activity levels of adults relative to juveniles; and the reduced ability of adults to utilize aerobic energy generation during exercise.

Given reductions in ATP synthesis of this magnitude, it is not surprising that another central feature of MYA is the reduced expression of active-transport processes, which are principal consumers of cellular ATP. Reduced expression of transporters may be highly relevant to the function of tissues, neuronal circuits, and organ systems. Reduced function of active transporters in tissues from middle-aged and aged adults has been observed in muscle, in neurons, and in kidney, and has been proposed to underlie declines in the functions of those tissues with age. For example, the Na⁺/K⁺ ATPase is used by neurons and muscle to restore membrane ion gradients following an action potential; this transporter may consume as much as 40% of cellular ATP. Reduced function of this transporter with age is seen in rat brain, and is proposed to contribute to the longer re-polarization times that are seen in neurons from aged brain (Fraser and Arief, 2001). Reduced expression of the sarco-endoplasmic reticulum Ca⁺ (SERCA) transporter is observed in aging in worms, flies, and mammals; adaptations to reduced SERCA expression are proposed to explain the prolonged action potentials that are characteristic of neurons and muscle from aged rodents (Janczewski et al., 2002; Pottorf et al., 2000). Active transport processes in the kidney also play a central role in regulating blood pressure and hydration by regulating osmotic balance in the blood; a blunted response of these transporters to dehydration is observed in rats as early as seven months of age and is proposed to underlie the kidneys' reduced ability to fight dehydration via urinary concentration (Amlal and Wilke, 2003).

18
22
UNIVERSIT
UNIVERSI
BI
UN
UNIVERSIT
UNIVERSI
BI
UN
UNIVERSIT
UNIVERSI
ALIFOR
ALIFOR
nci
AR
UNIVERSI
UNIVERS
18
UN

18
22
UNIVERSIT
UNIVERSI
BI
UN
UNIVERSIT
UNIVERSI
ALIFOR
ALIFOR
nci
AR
UNIVERSI
UNIVERS
18
UN



Model

Figure 4 presents one potential model for the regulation and effects of MYA. In this model, a signal from the developing reproductive system prevents cells from implementing MYA during development. After animals reach reproductive maturity, de-repression of MYA allows cells to implement changes that protect them from oxidative damage, at the potential expense of activity levels and intensely energy-consuming physiological processes.

MYA and human disease

The incidence rates of many human diseases increase quickly early in adulthood, while other diseases resolve at this time. One possible explanation is that the pattern of physiological changes in MYA increases vulnerability to some illnesses and reduces vulnerability to others. We consider a specific example, type II diabetes.

Type II diabetes is almost entirely an illness of adults. A salient feature of type II diabetes is reduced expression of mitochondrial oxidative metabolism (OXPHOS) genes in muscle, a primary target of insulin signaling (Mootha et al., 2003). Although this reduction in OXPHOS gene expression has not yet been causally implicated in type II diabetes, several lines of evidence suggest that quantitative perturbations of mitochondrial oxidative metabolism affect diabetes risk. Human type II diabetes is associated with mutations in two different mitochondrial-resident proteins, as well as with variation in PPAR-gamma, a regulator of mitochondrial gene expression (Altshuler et al., 2000; Florez et al., 2003). In mice, muscle-specific knockout of PPAR-gamma causes insulin resistance (Hevener et al., 2003). Aerobic exercise, a potent stimulator of OXPHOS gene expression, is highly protective against type II diabetes and

insulin resistance. If high levels of oxidative metabolism in muscle are indeed found to be protective against diabetes, then adult-onset reductions in OXPHOS gene expression may help explain why vulnerability to type II diabetes increases so steeply after humans reach adulthood.

MYA represents an exciting research direction

A research goal of high interest and relevance will be to identify the factors and molecular pathways that regulate the implementation of MYA. Two observations suggest that such a signal may originate in the reproductive system: (i) implementation of MYA commences soon after animals reach reproductive maturity; and (ii) signals from the reproductive system regulate aging and lifespan (Arantes-Oliveira et al., 2002; Hsin and Kenyon, 1999). Because MYA has many specific transcriptional targets, it will be possible to construct reporter strains that can be used to screen for mutations and RNAi treatments that affect the progression of these markers.

Another important goal will be to connect MYA to downstream, adult phenotypes. One straightforward approach in model organisms would be to artificially maintain high expression levels of genes whose expression is normally reduced by MYA. For example, does prolonging the high, juvenile-level expression of mitochondrial genes in adulthood keep animals youthful, active, and energetic (though perhaps at a cost to lifespan)? Does prolonging the adult expression of particular transporter genes keep certain physiological processes healthy (though perhaps at the expense of the energy available to other ones)? A mechanistic understanding of human familial hypodigoxinemic membrane Na(+)-K(+) ATPase regulatory syndrome, in which diverse

neurological and psychological pathologies co-exist with abnormally healthy aging and longevity (Kumar and Kurup, 2002), may also provide insight.

Understanding the physiological consequences of MYA will be helped by a better understanding of how cells respond to changes in energy availability – how a dwindling supply of energy is rationed across the various cellular and multicellular processes that allow cells to function in a multicellular organism.

A major emerging direction in aging research is to identify the timing of action of genes that affect aging and lifespan. Where it is found that the effect of a gene changes suddenly early in adulthood, as has been shown for insulin signaling and mitochondrial genes, that may suggest an interaction between MYA and the pathway of interest.

The conservation of MYA across large evolutionary distances suggests that model organisms will be useful in its elucidation, and that insights gained in *C. elegans*, *D. melanogaster*, and other model organisms will have high relevance to *H. sapiens*.

Conclusion

As a period of developmentally timed change that modulates essential relationships in aging and lifespan, young adulthood may offer an exciting connection between developmental biology and aging research. Understanding how animals change after they reach reproductive maturity may help us better understand our changing vulnerabilities to disease and the ancient tradeoffs that have shaped metazoan evolution.

ctis
RY
VERSIT
IVERSI
81
m
L
m
nci
RY
IVERSIT
IVERSI
-18
m
L
ALFORN
ALFORN
nci
ARC
NIVERSI
NIVERSI
181
m

181
m
L
ALFORN
ALFORN
nci
ARC
NIVERSI
NIVERSI
181
m

References

Altshuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M. C., Nemesh, J., Lane, C. R., Schaffner, S. F., Bolk, S., Brewer, C., *et al.* (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 26, 76-80.

Amlal, H., and Wilke, C. (2003). Resistance of mTAL Na⁺-dependent transporters and collecting duct aquaporins to dehydration in 7-month-old rats. *Kidney Int* 64, 544-554.

Arantes-Oliveira, N., Apfeld, J., Dillin, A., and Kenyon, C. (2002). Regulation of life-span by germ-line stem cells in *Caenorhabditis elegans*. *Science* 295, 502-505.

Blalock, E. M., Chen, K. C., Sharrow, K., Herman, J. P., Porter, N. M., Foster, T. C., and Landfield, P. W. (2003). Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment. *J Neurosci* 23, 3807-3819.

Bluher, M., Kahn, B. B., and Kahn, C. R. (2003). Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science* 299, 572-574.

Bluher, M., Michael, M. D., Peroni, O. D., Ueki, K., Carter, N., Kahn, B. B., and Kahn, C. R. (2002). Adipose tissue selective insulin receptor knockout protects against obesity and obesity-related glucose intolerance. *Dev Cell* 3, 25-38.

Clancy, D. J., Gems, D., Harshman, L. G., Oldham, S., Stocker, H., Hafen, E., Leivers, S. J., and Partridge, L. (2001). Extension of life-span by loss of CHICO, a *Drosophila* insulin receptor substrate protein. *Science* 292, 104-106.

Dillin, A., Hsu, A. L., Arantes-Oliveira, N., Lehrer-Graiwer, J., Hsin, H., Fraser, A. G., Kamath, R. S., Ahringer, J., and Kenyon, C. (2002). Rates of behavior and aging specified by mitochondrial function during development. *Science* 298, 2398-2401.

Felkai, S., Ewbank, J. J., Lemieux, J., Labbe, J. C., Brown, G. G., and Hekimi, S. (1999). CLK-1 controls respiration, behavior and aging in the nematode *Caenorhabditis elegans*. *Embo J* 18, 1783-1792.

Feng, J., Bussiere, F., and Hekimi, S. (2001). Mitochondrial electron transport is a key determinant of life span in *Caenorhabditis elegans*. *Dev Cell* 1, 633-644.

Florez, J. C., Hirschhorn, J., and Altshuler, D. (2003). The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu Rev Genomics Hum Genet* 4, 257-291.

Fraser, C. L., and Arief, A. I. (2001). Na-K-ATPase activity decreases with aging in female rat brain synaptosomes. *Am J Physiol Renal Physiol* 281, F674-678.

Hevener, A. L., He, W., Barak, Y., Le, J., Bandyopadhyay, G., Olson, P., Wilkes, J., Evans, R. M., and Olefsky, J. (2003). Muscle-specific Pparg deletion causes insulin resistance. *Nat Med* 9, 1491-1497.

Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G., and Brown, E. L. (2000). Genomic analysis of gene expression in *C. elegans*. *Science* 290, 809-812.

Hsin, H., and Kenyon, C. (1999). Signals from the reproductive system regulate the lifespan of *C. elegans*. *Nature* 399, 362-366.

Janczewski, A. M., Spurgeon, H. A., and Lakatta, E. G. (2002). Action potential prolongation in cardiac myocytes of old rats is an adaptation to sustain youthful intracellular Ca²⁺ regulation. *J Mol Cell Cardiol* 34, 641-648.

Jiang, C. H., Tsien, J. Z., Schultz, P. G., and Hu, Y. (2001). The effects of aging on gene expression in the hypothalamus and cortex of mice. *Proc Natl Acad Sci U S A* 98, 1930-1934.

Kayo, T., Allison, D. B., Weindruch, R., and Prolla, T. A. (2001). Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proc Natl Acad Sci U S A* 98, 5093-5098.

Kenyon, C., Chang, J., Gensch, E., Rudner, A., and Tabtiang, R. (1993). A *C. elegans* mutant that lives twice as long as wild type. *Nature* 366, 461-464.

Kimura, K. D., Tissenbaum, H. A., Liu, Y., and Ruvkun, G. (1997). *daf-2*, an insulin receptor-like gene that regulates longevity and diapause in *Caenorhabditis elegans*. *Science* 277, 942-946.

Kumar, A. R., and Kurup, P. A. (2002). Familial hypodigoxinemic membrane Na(+)-K(+) ATPase upregulatory syndrome - relation between digoxin status and cerebral dominance. *Neurol India* 50, 340-347.

Lee, C. K., Klopp, R. G., Weindruch, R., and Prolla, T. A. (1999). Gene expression profile of aging and its retardation by caloric restriction. *Science* 285, 1390-1393.

Lee, C. K., Weindruch, R., and Prolla, T. A. (2000). Gene-expression profile of the ageing brain in mice. *Nat Genet* 25, 294-297.

Lee, S. S., Lee, R. Y., Fraser, A. G., Kamath, R. S., Ahringer, J., and Ruvkun, G. (2003). A systematic RNAi screen identifies a critical role for mitochondria in *C. elegans* longevity. *Nat Genet* 33, 40-48.

Lin, K., Hsin, H., Libina, N., and Kenyon, C. (2001). Regulation of the *Caenorhabditis elegans* longevity protein DAF-16 by insulin/IGF-1 and germline signaling. *Nat Genet* 28, 139-145.

Lund, J., Tedesco, P., Duke, K., Wang, J., Kim, S. K., and Johnson, T. E. (2002). Transcriptional profile of aging in *C. elegans*. *Curr Biol* 12, 1566-1573.

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C. S., Jan, Y. N., Kenyon, C., Bargmann, C. I., and Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet* 36, 197-204.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al.* (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34, 267-273.

Murphy, C. T., McCarroll, S. A., Bargmann, C. I., Fraser, A., Kamath, R. S., Ahringer, J., Li, H., and Kenyon, C. (2003). Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* 424, 277-283.

Pletcher, S. D., Macdonald, S. J., Marguerie, R., Certa, U., Stearns, S. C., Goldstein, D. B., and Partridge, L. (2002). Genome-wide transcript profiles in aging and calorically restricted *Drosophila melanogaster*. *Curr Biol* 12, 712-723.

Pottorf, W. J., Duckles, S. P., and Buchholz, J. N. (2000). SERCA function declines with age in adrenergic nerves from the superior cervical ganglion. *J Auton Pharmacol* 20, 281-290.

Tatar, M., Kopelman, A., Epstein, D., Tu, M. P., Yin, C. M., and Garofalo, R. S. (2001). A mutant *Drosophila* insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science* 292, 107-110.

Weindruch, R., Kayo, T., Lee, C. K., and Prolla, T. A. (2001). Microarray profiling of gene expression in aging and its alteration by caloric restriction in mice. *J Nutr* 131, 918S-923S.

Welle, S., Brooks, A. I., Delehanty, J. M., Needler, N., Bhatt, K., Shah, B., and Thornton, C. A. (2004). Skeletal muscle gene expression profiles in 20-29 year old and 65-71 year old women. *Exp Gerontol* 39, 369-377.

Williams, G. C. (1957). Pleiotropy, natural selection and the evolution of senescence. *Evolution* 11, 398-411.

Table 1

Conservation between humans and invertebrates of patterns of gene expression change observed during the progression of adulthood. All profiles are compared to profile from Welle et al., 2004, in which gene expression in *vastus lateralis* muscle from young adult (20-29 years) and aged adult (65-71 years) human females is compared.

Organism	Samples compared	Reference	R	Gene pairs	P
<i>C. elegans</i>	144 hrs. adult age vs. 0 hrs. adult age	McCarroll et al., 2004	0.117	1160	$P < 10^{-5}$
	2 wks. age vs. 4 days age	Hill et al., 2000	0.256	256	$P < 10^{-5}$
<i>D. melanogaster</i>	adult d18 vs. adult d3, whole females	Pletcher et al., 2001	0.152	1107	$P < 10^{-7}$
	adult d47 vs. adult d3, males, heads	McCarroll et al., 2004	0.187	999	$P < 10^{-8}$

Figure 1

Adult-onset changes in gene expression in *H. sapiens* skeletal muscle (data from Welle et al., 2004) and *D. melanogaster* heads (data from McCarroll et al., 2004). Each small box shows the adult-onset change in expression of a single gene. Orthologous genes from the two organisms are stacked vertically. Each large rectangle corresponds to genes in a Gene Ontology category identified using the approach published earlier (McCarroll et al., 2004).

BIOLOGICAL PROCESSES

Oxidative phosphorylation (GO:0006119, n = 24, $p < 10^{-4}$)



Main pathways of carbohydrate metabolism (GO:0006092, n = 19, $p < 10^{-4}$)



Tricarboxylic acid cycle (GO:0006099, n = 13, $p < 10^{-4}$)



Ion transport (GO:0006811, n = 18, $p < 10^{-10}$)



CELLULAR COMPONENTS

Mitochondrial matrix (GO:0005759, n = 53, $p < 10^{-3}$)



Mitochondrial inner membrane (GO:0005743, n = 46, $p < 10^{-9}$)



MOLECULAR PATHWAYS

DNA binding (GO:0003677, n = 40, $p < 10^{-3}$)



Oxidoreductase activity (GO:0016491, n = 80, $p < 10^{-6}$)



P-P-bond-hydrolysis-driven transporter activity (GO:0015405, n = 28, $p < 10^{-6}$)



ATPase activity, coupled to transmembrane movement of substances (GO:0042626, n = 25, $p < 10^{-7}$)



Monovalent inorganic cation transporter activity (GO:0015077, n = 30, $p < 10^{-7}$)



Primary active transporter activity (GO:0015399, n = 45, $p < 10^{-10}$)



Figure 2

Some of the conserved patterns of adult-onset gene expression change shared by *H. sapiens* skeletal muscle (Welle et al., 2004), *D. melanogaster* heads (McCarroll et al., 2004), and *C. elegans* (McCarroll et al., 2004).

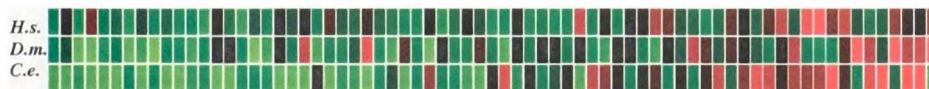
Mitochondrial inner membrane (GO:0005759)



Primary active transporter activity (GO: 0015399)



Oxidoreductase activity (GO:0016491)



Carbohydrate metabolism (GO:0006092)



Figure 3

Timing of repression of oxidative metabolism genes in *C. elegans* (data from McCarroll et al., 2004), *D. melanogaster* (data from Pletcher et al., 2002), and *R. norvegicus* hippocampal neurons (data from Blalock et al., 2003). Plotted points show median expression level change (relative to baseline) of oxidative metabolism genes; error bars show standard error around the mean expression level change.

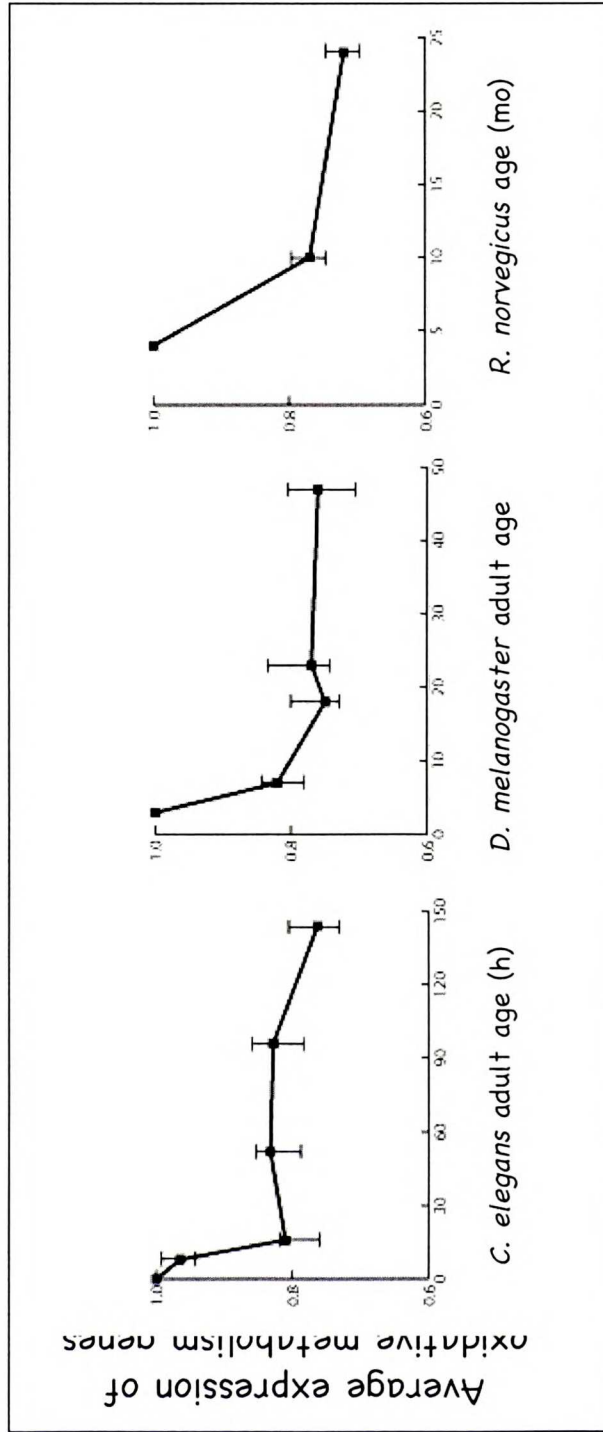


Figure 4

A model for the developmental regulation and adult effects of the MYA (metazoan young adult) gene expression program.

During development:

Developing reproductive system — MYA gene expression program

High oxidative metabolism, ROS generation (lifespan-limiting)
High ATP utilization
High activity, physiological function
daf-2/daf-16 regulate dauer formation but not lifespan

During adulthood:

Mature reproductive system — MYA gene expression program

Reduced oxidative metabolism, ROS generation (less lifespan-limiting)
Reduced ATP utilization
Reduced function of transport processes
Reduced physiological functions
Altered insulin responsiveness
daf-2/daf-16 regulate lifespan

UNIVERSITY OF CALIFORNIA
LIBRARY

UNIVERSITY OF CALIFORNIA
LIBRARY

CHAPTER 3
**DECODING THE TRANSCRIPTIONAL REGULATION OF
CHEMOSENSORY RECEPTOR GENES IN *C. ELEGANS***

Abstract

Understanding how promoter sequences encode gene regulation is a major goal in biology. Large families of related genes, such as chemosensory receptors, may offer unique opportunities for decoding promoter sequences. Here we use a statistical approach, probabilistic segmentation, to identify a large number of candidate transcriptional control sequences or motifs in the promoters of the *C. elegans* chemosensory receptor genes. Many of these motifs show positional preference, are specific to chemosensory receptor genes, and correspond to the binding sites of known families of transcription factors in different organisms. We have functionally characterized one of these motifs, the E-box sequence WWYCACSTGY, and found that it confers expression in the ADL chemosensory neurons.

Introduction

The transcriptional regulation of a gene is determined by the non-protein-coding sequences around and within it, most frequently by its upstream, promoter sequence. While scientists have long understood how the protein-coding parts of genes encode protein sequences, comparatively little is understood about how regulatory sequence determines the place, time, and manner of gene expression. The advent of genome-scale sequence data for many

organisms offers new opportunities for decoding regulatory sequences. In particular, the sequences of large families of functionally related genes, which are likely to share many regulatory motifs, may offer new opportunities for finding such transcriptional control sequences.

The *C. elegans* genome encodes more than one thousand candidate chemosensory receptors -- G-protein coupled receptors with potential roles in olfaction and taste. The expression patterns of dozens of these chemosensory receptors have been assayed using promoter-gfp fusions; they show diverse spatial expression patterns, but each receptor is typically expressed in one or a few chemosensory neuron pairs (Troemel et al., 1995). The promoters of many chemosensory receptor genes also appear to be regulated by starvation, by sensory activity, and by dauer formation and recovery (Nolan et al., 2002; Peckol et al., 2001).

The transcriptional regulation of chemosensory receptors is of particular biological interest because it has the potential to program and reprogram chemosensory behavior -- during development, in response to environmental changes, and throughout evolution. In *C. elegans*, the G proteins and their downstream effectors are widely expressed in sensory neurons, allowing chemosensory receptors to be behaviorally active when expressed ectopically in other neurons (Troemel et al., 1997). For example, the animal's attractive chemotactic response to the volatile odorant diacetyl is mediated by its receptor ODR-10, which is expressed in the AWA chemosensory neurons. Chemotaxis to diacetyl is restored in an *odr-10* mutant by expressing ODR-10 in the AWC chemosensory neurons, which sense other attractive stimuli (Wes and Bargmann, 2001). The behavioral response to diacetyl is reprogrammed into an avoidance

response by mis-expressing ODR-10 in the AWB chemosensory neurons, which sense aversive stimuli (Troemel et al., 1997). Thus, *C. elegans* could use transcriptional regulation to reprogram its mapping of environmental cues onto behavioral responses. The promoter sequences of the chemosensory receptor genes match the animal's specific behavioral responses with specific chemical cues in its environment.

Genetic studies of chemosensation and chemosensory cell fate have identified several transcription factors that are required for appropriate cell-specific receptor expression (Chang et al., 2003; Colosimo et al., 2003; Lanjuin and Sengupta, 2004; Lanjuin et al., 2003; Sagasti et al., 1999; Sarafi-Reinach et al., 2001; Sarafi-Reinach and Sengupta, 2000; Satterlee et al., 2001; Sengupta et al., 1994). Relatively little is known about how the promoter sequences of chemosensory receptor genes encode their regulation.

Regulatory elements in promoter sequences have traditionally been identified by experimental analysis of the effects of deletions of promoter sequence upon reporter gene transcription. This approach provides definitive experimental evidence of the function of an element, but it requires extensive experimental work for each characterized promoter. With the advent of genome-scale information, two approaches to identifying regulatory elements have shown promise. First is the use of microarray data to identify co-regulated sets of genes. In this approach, DNA microarrays are used to identify sets of genes that show correlated patterns of expression across many experiments; the promoters from these gene sets are then searched for shared sequence motifs (Bussemaker et al., 2001; Gaudet and Mango, 2002; Patil et al., 2004). A requirement of this approach is that genes must be expressed at a sufficiently

high level to yield reliable expression measurements; this requirement is not met by the chemoreceptor genes, most of which are expressed in only a few cells and are not detectable on microarrays. A second approach to finding regulatory elements involves identifying conserved, non-coding sequences in the genomes of related organisms (Kellis et al., 2003; Loots et al., 2000; Zheng et al., 2004). The critical assumption of this approach is that orthologous genes in the two organisms are actually regulated in the same way in both organisms. This requirement is not met by the *Caenorhabditis* chemoreceptor gene family, whose rapid evolution appears to have led to non-conserved expression patterns in *C. elegans* and *C. briggsae*. In this paper we describe a third approach, which assumes neither high expression levels nor homologous expression patterns: the statistical analysis of sequence motifs in sets of genes of related function.

The approach we employ here utilizes the strengths of the chemosensory gene family in *C. elegans*: the large number of genes it encompasses, and the appearance of shared regulatory themes across large subsets of the gene family. These properties make the chemosensory receptor gene set appropriate for probabilistic segmentation, which can find regulatory motifs in a set of genes whose regulation is heterogeneous but shows common themes (Bussemaker et al., 2000). This method is based on the identification of short DNA sequences, or “words”, that are statistically over-represented in a set of sequences. Probabilistic segmentation makes efficient use of information that is dispersed across a large number of genes, while making minimal assumptions about how regulatory elements are distributed across those genes. Importantly, a sequence can be statistically over-represented even if it is present in only a small fraction of the gene promoters. The approach seems ideally suited to the dissection of *C.*

C. elegans chemosensory receptor gene promoters, in which we expect that large numbers of genes share common regulation, but have little *a priori* information about which specific genes are co-regulated. For example, we expect that subsets of the *C. elegans* chemoreceptor genes will be expressed in the same cells, or will be regulated by activity or by dauer pheromone; however, we do not know *a priori* which genes will share these regulatory themes.

We use this approach to identify a set of candidate regulatory elements in *C. elegans* chemosensory receptor promoters, and show that one of these elements is a novel chemosensory regulatory motif.

RESULTS

Segmentation of *C. elegans* chemoreceptor promoter sequences into motifs

To create a sequence data set for analysis, we used the predicted structures of 921 likely chemosensory receptor genes of the *sra*, *srb*, *src*, *srd*, *sre*, *srh*, *sri*, *srj*, *srm*, *srn*, *sro*, *srp*, *srr*, *srs*, *sru*, *srv*, *srw*, *srx*, and *str* families, as predicted by Hugh Robertson (<http://www.wormbase.org>). We extracted 1 kb of sequence upstream of the predicted translational start site of each of these predicted genes, removing those sequences that intersected the coding sequences of other genes. Repetitive sequence was filtered from this data set using the REPUTER algorithm (Kurtz and Schleiermacher, 1999).

The MobyDick probabilistic segmentation algorithm (Bussemaker et al., 2000) was applied to this sequence data set to build a dictionary for the putative chemosensory receptor promoter DNA. About 8% of the sequence was

segmented into some 400 words of length 6 or greater, which collectively appeared 7,345 times, or about 20 times each.

Positional preference of candidate motifs

To explore the positional preference of candidate motifs, all occurrences of each motif in the upstream sequences of the chemosensory receptor genes were identified, and the positions of these appearances, relative to the predicted translational start site, were recorded. This resulted in a distribution of positions for each motif, which were divided into positional regions or “bins” (-999 to -900; -899 to -800; ... -99 to ATG). The Chi-squared test was used to assess whether the distribution of occurrences across these bins differed significantly from a uniform distribution.

Twelve of the candidate motifs met a statistical cutoff of $p < 10^{-4}$ ($p < 0.1$ after Bonferroni correction). These motifs are listed in Table 2. All twelve motifs showed strong preference for the proximal promoter region, tending to occur within 200 nt of the predicted translational start site (ATG). All twelve sequences are binding sites for known families of transcription factors:

Motifs with an E-box core. Nine of these twelve motifs shared the E-box core sequence CASCTG on either the coding or non-coding strand. E-boxes (CANNTG) are bound by transcription factors of the basic helix-loop-helix family; specificity for particular family members is determined by the two interior nucleotides (NN) and by the nucleotides flanking the E-box core. The frequencies of the core E-box sequences CACCTG, CAGGTG, and CAGCTG in *C. elegans* chemoreceptor promoters all peaked between -40 and -120 (Figure 1A). By contrast, the similar E-box sequence CACGTG (which did not appear in the

ASIT
RSI
n
n
RSIT
RSI
n
n
ALFO
ALFO
aci
R
IVERSI
IVERSI
IB
n

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

probabilistic segmentation results) did not show any positional preference within the chemoreceptor gene family (Figure 1A).

SMAD-binding motifs. Another two motifs, GTCTAG and CTAGAC, are complementary sequences with common positional preference, suggesting that the motif they identify can appear on either the coding or the non-coding strand. The frequency of these motifs is greatest at positions between -40 and -180 (Figure 1B). In mice, this sequence is bound by transcription factors of the SMAD family (Norwitz et al., 2002).

CdxA-binding sequence. The CTATAATT motif showed a positional preference that peaked between -60 and -120; the motif also showed a strand preference, with more appearances at almost every position than its reverse complement sequence AATTATAG, which had not been identified in the probabilistic segmentation (Figure 1C). CTATAATT has been identified experimentally as a binding site for the mammalian caudal-type homeobox domain transcription factor CdxA (Margalit et al., 1993), whose orthologs in *C. elegans* include *pal-1*, *lin-39*, and *ceh-13*.

Functional preference of candidate transcriptional motifs

If a sequence functions in a particular transcriptional pathway, then the genes containing this motif in their upstream sequences should tend to have a common molecular or biological function. To assess the functional specificity of candidate transcriptional motifs, we first identified all occurrences of that motif in the proximal promoters of all predicted *C. elegans* genes. We then used the Gene Ontology (GO) categories (Ashburner et al., 2000) to identify about 600 sets of genes defined by common molecular function, by localization of their

UNIVERSITY OF CALIFORNIA
LIBRARY

UNIVERSITY OF CALIFORNIA
LIBRARY

products to a common cellular component, or by a common biological role. For each combination of motif and GO category, we assessed whether the motif was over-represented in that GO category, relative to its frequency across all other genes. All significant results for the three sequence motifs with strong positional preference (the E-box core, the SMAD-binding site, and the CdxA-binding site) are listed in Table 3.

The E-box core sequences cacctg (on either strand) and cagctg were overrepresented in the proximal promoter regions of two functional sets of genes: G-protein coupled receptors and histones (Table 3). (By contrast, the E-box motif cagctg, which had not been identified by probabilistic segmentation analysis, was overrepresented in cell motility genes, due to its appearances in the proximal promoters of most genes with major sperm protein (MSP) domains; we suggest that the cagctg motif may represent a promoter element used to drive gene expression in sperm.) The candidate SMAD-binding motif and the candidate CdxA motif were both over-represented in the proximal promoters of G-protein coupled receptors, but not of other gene categories. These three motifs thus appear to show high functional specificity.

Distribution of motifs across chemoreceptor subfamilies

The *C. elegans* chemoreceptor genes have been classified into multiple families based upon their coding sequences (Robertson, 1998; Robertson, 2000). Although only one of these genes has a known ligand, the diacetyl receptor ODR-10 (Sengupta et al., 1996), the chemoreceptor families may well correspond to families of related ligands, so we sought to ascertain whether the promoter sequence elements were overrepresented in particular chemoreceptor families.

The E-box sequences were strongly over-represented in the *srh* and *sri* families: they appeared in the proximal promoter of 36 of 59 *sri* genes (61% versus a genome-wide frequency of 4.4%, $p < 10^{-34}$) and 100 of 222 *srh* genes (45% versus a genome-wide frequency of 4.4%, $p < 10^{-74}$) but were not over-represented in the proximal promoters of other chemoreceptor subfamilies (Figure 2). This was not due simply to recent duplication of genes in the *srh* and *sri* families: the E-box sequences were far more broadly shared than any other sixmer in the promoters in the *srh* and *sri* gene families, and within these families, there was no significant relationship between the homology of coding sequences and the likelihood that promoter E-box sequences were shared. The SMAD motif was over-represented in genes of the *str* family, appearing in 26 of 190 genes (14% versus a background frequency in the genome of 3.2%; $p < 10^{-9}$). The CdxA motif was distributed across chemoreceptor subfamilies in a way that was not significantly different from a random distribution.

Sequence context of candidate transcriptional motifs

To assess the immediate sequence context in which candidate transcriptional motifs occur, we extracted the flanking sequence around each appearance of each of these motifs in chemosensory receptor genes, and analyzed the distribution of nucleotide frequencies at flanking sites. The significance of the nucleotide distributions at flanking sites was assessed in two ways. First, the distribution of nucleotide frequencies at a site was compared to the background nucleotide frequencies of *C. elegans* upstream sequences using the Chi-square distribution. Second, the “information” present at a nucleotide position was calculated; high information content suggests that nucleotide

UNIVERSITY OF CALIFORNIA LIBRARY

UNIVERSITY OF CALIFORNIA LIBRARY

frequencies at that position are biased by the presence of the core site (Schneider et al., 1986). These two approaches reached similar results.

Appearance of the asymmetric e-box core sequence CACCTG on either DNA strand in chemoreceptor promoter sequences strongly biased the nucleotide composition of flanking sites (Table 5), suggesting the larger motif WWYCACCTGY. We obtained the same consensus sequence when we separately examined occurrences of CACCTG on the coding and non-coding strands. This WWYCACCTGY motif describes seven of the highest-scoring words identified in the original probabilistic segmentation and many lower-scoring words as well. When a scoring matrix developed from these 143 sites was used to orient the symmetric motif CAGCTG (by choosing the higher-scoring orientation), the distribution of the resulting scores was not significantly different from that of the asymmetric sites, and it significantly exceeded that of a control distribution of scores generated from random flanking sequences. Thus, the larger motif WWYCASCTGY appears to describe a consensus sequence for the occurrences of the E-box CASCTG in the proximal promoters of chemoreceptor genes.

Because CASCTG E-boxes are also enriched in the proximal promoters of *C. elegans* histone genes (Table 3), we analyzed the flanking sequences surrounding the occurrences of CASCTG in the proximal promoters of 29 histone genes, and obtained the consensus sequence CAYSRCASSTG, which differed strongly from the chemoreceptor consensus. We also analyzed the flanking sequences surrounding the occurrences of CACGTG in the proximal promoters of 17 major sperm protein genes, and obtained the symmetric consensus sequence TYCACGTGRA, which differed from both the chemoreceptor and

histone consensus sequences. The strong, distinct consensus sequences for the E-boxes in these three gene families, with differences at positions known to determine the specificity of transcription factor binding, suggest that they are bound by distinct transcription factors.

Sequences surrounding appearances of the CdxA motif and the SMAD motif did not differ significantly in nucleotide frequencies from promoter sequences in general.

The WWYCASCTGYG motif appears in ADL-expressed genes

To better understand the biological relevance of our computational results, we sought to functionally characterize the E-box motif WWYCASCTGYG. We began by looking for chemosensory receptor genes with proximal E-boxes and known expression patterns. The expression patterns of many candidate chemosensory receptor genes in *C. elegans* have been assessed by promoter::GFP fusions (Troemel et al., 1995). We searched the promoter sequences of these genes for proximal E-boxes. The two genes with the strongest matches to the full consensus sequence WWYCASCTGYG in their proximal promoters were both expressed in the ADL chemosensory neuron pair. The *srh-220* gene, which has the sequence tttcacctgct at a position 96 nucleotides upstream of its ATG, is expressed in ADL. The *sro-1* gene, which has the sequence ttccagctggt at a position 66 nucleotides upstream of its ATG, is expressed in ADL and SIA. Several other chemosensory receptor genes have proximal E-boxes for which the surrounding sequence does not match the full consensus motif; these genes show distinct, generally non-overlapping expression patterns (Table 6, below line).

To assess whether genes with proximal WWYCASCTGY motifs are expressed in ADL more generally, the promoters of additional chemosensory receptor genes were cloned into an expression vector for the green fluorescent protein (GFP) and used to establish transgenic lines. The promoters of genes with strong matches to the consensus sequence all drove robust expression in ADL. An *srh-186::GFP* transgene showed robust GFP expression in ADL in all three lines examined. An *sri-51::GFP* transgene showed GFP expression in ADL, PHA, PHB, and a head interneuron in all three lines examined. An *srh-132::GFP* transgene showed GFP expression in ADL and non-neuronal tissues in the central body, in all three lines examined.

The WWYCACSTGY motif functions as an ADL enhancer

To assess whether the consensus WWYCACSTGY sequence drives expression in ADL, we inserted this sequence (ttcacctgtt) into the proximal promoter of the *odr-10* chemosensory receptor gene, which is expressed in the AWA neuron. Animals carrying a *Podr-10(+ttcacctgtt)::gfp* transgene showed bright GFP expression in ADL (Table 6, Figure 3). This result suggests that this 11-nucleotide sequence is capable of driving expression in the ADL neurons.

Identifying transcriptional regulators of the ADL-E-box pathway

To increase our understanding of the transcriptional pathway implemented through the WWYCACSTGY motif, we sought to identify candidate transcription factors in the ADL gene expression pathway. E-boxes are bound by the basic-helix-loop-helix (bHLH) family of transcription factors, 31 of which are encoded by the *C. elegans* genome.

UNIVERSITY OF ALABAMA
LIBRARY

UNIVERSITY OF ALABAMA
LIBRARY

The physical structure of bHLH binding to DNA has been extensively characterized via structural and mutational analyses. The specificity of bHLH transcription factors for sequences within the E-box core is encoded in the thirteenth residue of the basic domain, which contacts the E-box core in the crystal structure of the bHLH protein Pho4p (Shimizu et al., 1997). A mutant form of MyoD at this residue (L13 to R13) recognizes a c-Myc binding site (CACGTG) instead of a MyoD binding site (CAGCTG) (Blackwell et al., 1993). Additional information is provided by the flanking sequence of the ADL motif. A thymidine nucleotide 5' to the E-box core, as frequently found in the WWYCACSTGY motif, inhibits the binding of Pho4p, due to the presence of a Glu residue in the third residue of the basic region of Pho4p (Fisher and Goding, 1992). When this Glu is mutated to the Cbf1p counterpart (Asp), the mutant Pho4p can recognize the core sequence flanked by the thymidine nucleotide (Fisher and Goding, 1992). Only four of the *C. elegans* bHLH transcription factors have both a large, non-polar residue at position 13 and a small, polar residue at position 3 of the basic domain; these are *hlh-2*, *lin-32*, *ngn-1*, and *hlh-12*.

The *hlh-2* gene is of particular interest. *hlh-2* is expressed in all nuclei of the early embryo for the first 150-200 minutes of development, and then is restricted to 21 cells, including ADL, ASH, RIC, and various cells in the tail and intestine (Krause et al., 1997). *hlh-2* encodes a homolog of the *Drosophila* *daughterless* gene, which is required for formation of peripheral neurons and associated sensory structures (Caudy et al., 1988a; Caudy et al., 1988b; Murre et al., 1989). *In vitro*, HLH-2 heterodimerizes with LIN-32 and binds to an E-box from the *hlh-2* promoter (tttcacctgct) that is a perfect match to the consensus sequence identified here; HLH-2 may also weakly bind this sequence as a

18
18
UNIVERSITY
UNIVERSITY
18
Law
L
UNIVERSITY
Francis
UNIVERSITY
UNIVERSITY
18
Law
L
UNIVERSITY
UNIVERSITY
Francis
UNIVERSITY
UNIVERSITY
18
Law

Francis
UNIVERSITY
UNIVERSITY
18
Law
L
UNIVERSITY
UNIVERSITY
Francis
UNIVERSITY
UNIVERSITY
18
Law

homodimer (Portman and Emmons, 2000). Two weak missense *hlh-2* alleles exist; these alleles show no reported phenotype on their own (Portman and Emmons, 2000), and indeed we found that ADL expression of *sri-51::gfp* and *odr-10(+ADLmotif)::gfp* persisted in *hlh-2(bx108)* and *hlh-2(bx115)* backgrounds. The *lin-32* gene is transiently expressed in cells of the ray sublineage, with no reported expression in head neurons (Portman and Emmons, 2000). Nonetheless, to test the hypothesis that expression in ADL via the TTYCASCTGY motif might require LIN-32, we crossed *sri-51::gfp* and *odr-10(+ADLmotif)::gfp* into a *lin-32(u282)* background. ADL expression of both transgenes persisted in *lin-32(u282)* mutants, though the expression of *sri-51::gfp* in interneurons appeared to be extinguished in the *lin-32(u282)* background. We speculate that, in ADL, HLH-2 may heterodimerize with a different bHLH family member that is expressed in ADL. *lin-32* is most similar to *cnd-1*, *ngn-1*, and *hlh-10*. The *hlh-12* and *hlh-10* genes have not been characterized. The *ngn-1* gene has been characterized by RNAi, which causes defects of axon guidance in the nerve ring, and by promoter::GFP fusion, which indicates embryonic expression in about 10 neurons until the three-fold stage.

Discussion

We used a computational approach to identify candidate transcriptional control elements in the *C. elegans* chemoreceptor gene family. Our approach did not require microarray data or genome sequences from other nematodes, and could therefore be applied to organisms and gene families for which such resources are not currently available. The approach identified numerous candidate sequences with characteristics of transcriptional control elements:

positional preference; preference for functionally defined sets of genes; and sequences that are bound by known families of transcription factors. We functionally characterized one of these motifs and found it to be a potent enhancer that drives expression specifically in the ADL chemosensory neurons.

We found that the insertion of a short, 11-nucleotide sequence into the promoter of a *C. elegans* chemosensory receptor gene causes that gene to be expressed in ADL. This suggests that very short enhancer sequences can contribute to the expression of other genes in a modular way. This result suggests a potential model for the evolution of *C. elegans* chemosensory behaviors: the appearance or mutation of small sequences in chemosensory receptor promoters could re-program behavioral responses to particular odors by adding or removing expression in particular chemosensory neurons.

The ADL enhancer element identified in this paper is present in approximately half of the promoters of chemoreceptor genes in the *srh* and *sri* chemoreceptor subfamilies. This is the first broad association of a particular chemoreceptor subfamily with a particular chemosensory neuron. This relationship might reflect the use of ADL to sense a particular class of ligands. Understanding this, however, will require considerable progress: only one receptor-ligand relationship is known in *C. elegans* (Sengupta et al., 1996), and almost nothing is known about chemicals sensed by ADL.

Other candidate motifs identified by our approach seem likely to be functional transcriptional control elements as well. For example, the CTAGAC motif consists of a sequence that is bound by transcription factors of the SMAD family in mice (Norwitz et al., 2002). *C. elegans* SMADs are implicated in the regulation of dauer formation (Daniels et al., 2000; Inoue and Thomas, 2000).

Dauer formation regulates the expression of the *str-2* and *str-3* chemosensory receptor genes (Peckol et al., 2001), both of which have this motif in their proximal promoters.

The identification of a sequence motif as a high-quality “word” implies that that sequence corresponds to a meaningful unit of biological information; but it does not necessarily implicate a sequence in transcriptional regulation. Sequence motifs might play any of a number of other biological roles, such as contributing to secondary structure, interacting with chromatin, or participating in recombination. Several features might distinguish transcriptional control elements from other elements. One such feature is positional preference – a tendency to appear at particular physical positions relative to genes – which reflects the physical mechanism by which the binding of transcription factors influences transcription. Another feature is functional preference -- a tendency to appear in front of genes that have common biological roles – which reflects the broader regulatory logic of transcriptional control systems. Many potentially promising motifs were unable to meet the demanding statistical criteria applied to assess positional preference here, because the requirement of high significance after Bonferroni correction limited analysis to common motifs and motifs with extreme positional preference. These criteria excluded less-common motifs with moderate but biologically significant positional preference. Although we have focused on the motifs in Table 2, this result should in no way limit the potential biological relevance of the other words in the dictionary.

For researchers who seek to identify likely control sequences in promoters of interest, or to identify candidate genes that may share expression patterns

with genes of interest, the candidate motifs in Table 2 may prove a useful starting point.

Methods

Sequence data. To create a sequence data set for analysis, we used the predicted structures of 921 likely chemosensory receptor genes of the *sra*, *srb*, *src*, *srd*, *sre*, *srh*, *sri*, *srj*, *srn*, *sro*, *srp*, *srr*, *srs*, *sru*, *srv*, *srw*, *srx*, and *str* families, as predicted by Hugh Robertson (<http://www.wormbase.org>).

Probabilistic segmentation. We used the MobyDick algorithm (Bussemaker et al., 2000) to build a dictionary for the *C. elegans* chemosensory receptor promoters. Starting from a dictionary of one-letter words ("a", "c", "t", "g"), words were extended via an exhaustive search procedure and retained when their inclusion contributed to $P(S|D)$, as described elsewhere (Bussemaker et al., 2000). This procedure was repeated until the results converged, i.e. until additional iterations did not change the dictionary. The final dictionary had 589 words, of lengths 1 to 11 bp. On average, 58% of the sequence data were segmented into single-letter words and 92% into words of length five or less (Table 1); this distribution was broadly similar to an earlier probabilistic segmentation of promoter sequences from *S. cerevisiae* (Bussemaker et al., 2000).

GFP transgenes. PCR was used to amplify fragments of 3-4 kb of genomic data upstream of the *srh-186*, *sri-51*, *sri-132*, and *srw-95* genes, and to introduce flanking FseI and AscI restriction sites. These fragments were then directionally cloned into the FseI and AscI sites of the *pSM/gfp* expression vector. Transgenes were injected (100 ng/ul) with an *odr-1::dsRed* co-injection marker (40 ng/ul).

References

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29.
- Blackwell, T. K., Huang, J., Ma, A., Kretzner, L., Alt, F. W., Eisenman, R. N., and Weintraub, H. (1993). Binding of myc proteins to canonical and noncanonical DNA sequences. *Mol Cell Biol* 13, 5216-5224.
- Bussemaker, H. J., Li, H., and Siggia, E. D. (2000). Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A* 97, 10096-10100.
- Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nat Genet* 27, 167-171.
- Caudy, M., Grell, E. H., Dambly-Chaudiere, C., Ghysen, A., Jan, L. Y., and Jan, Y. N. (1988a). The maternal sex determination gene daughterless has zygotic activity necessary for the formation of peripheral neurons in *Drosophila*. *Genes Dev* 2, 843-852.

Caudy, M., Vassin, H., Brand, M., Tuma, R., Jan, L. Y., and Jan, Y. N. (1988b). daughterless, a Drosophila gene essential for both neurogenesis and sex determination, has sequence similarities to myc and the achaete-scute complex. *Cell* 55, 1061-1067.

Chang, S., Johnston, R. J., Jr., and Hobert, O. (2003). A transcriptional regulatory cascade that controls left/right asymmetry in chemosensory neurons of *C. elegans*. *Genes Dev* 17, 2123-2137.

Colosimo, M. E., Tran, S., and Sengupta, P. (2003). The divergent orphan nuclear receptor ODR-7 regulates olfactory neuron gene expression via multiple mechanisms in *Caenorhabditis elegans*. *Genetics* 165, 1779-1791.

Daniels, S. A., Ailion, M., Thomas, J. H., and Sengupta, P. (2000). *egl-4* acts through a transforming growth factor-beta/SMAD pathway in *Caenorhabditis elegans* to regulate multiple neuronal circuits in response to sensory cues. *Genetics* 156, 123-141.

Fisher, F., and Goding, C. R. (1992). Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. *Embo J* 11, 4103-4109.

Gaudet, J., and Mango, S. E. (2002). Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science* 295, 821-825.

Inoue, T., and Thomas, J. H. (2000). Targets of TGF-beta signaling in *Caenorhabditis elegans* dauer formation. *Dev Biol* 217, 192-204.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241-254.

Krause, M., Park, M., Zhang, J. M., Yuan, J., Harfe, B., Xu, S. Q., Greenwald, I., Cole, M., Paterson, B., and Fire, A. (1997). A *C. elegans* E/Daughterless bHLH protein marks neuronal but not striated muscle development. *Development* 124, 2179-2189.

Kurtz, S., and Schleiermacher, C. (1999). REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* 15, 426-427.

Lanjuin, A., and Sengupta, P. (2004). Specification of chemosensory neuron subtype identities in *Caenorhabditis elegans*. *Curr Opin Neurobiol* 14, 22-30.

Lanjuin, A., VanHoven, M. K., Bargmann, C. I., Thompson, J. K., and Sengupta, P. (2003). Otx/otd homeobox genes specify distinct sensory neuron identities in *C. elegans*. *Dev Cell* 5, 621-633.

Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. (2000). Identification of a coordinate regulator of

interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288, 136-140.

Margalit, Y., Yarus, S., Shapira, E., Gruenbaum, Y., and Fainsod, A. (1993). Isolation and characterization of target sequences of the chicken CdxA homeobox gene. *Nucleic Acids Res* 21, 4915-4922.

Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L. Y., Jan, Y. N., Cabrera, C. V., Buskin, J. N., Hauschka, S. D., Lassar, A. B., and et al. (1989). Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell* 58, 537-544.

Nolan, K. M., Sarafi-Reinach, T. R., Horne, J. G., Saffer, A. M., and Sengupta, P. (2002). The DAF-7 TGF-beta signaling pathway regulates chemosensory receptor gene expression in *C. elegans*. *Genes Dev* 16, 3061-3073.

Norwitz, E. R., Xu, S., Jeong, K. H., Bedecarrats, G. Y., Winebrenner, L. D., Chin, W. W., and Kaiser, U. B. (2002). Activin A augments GnRH-mediated transcriptional activation of the mouse GnRH receptor gene. *Endocrinology* 143, 985-997.

Patil, C. K., Li, H., and Walter, P. (2004). Gcn4p and novel upstream activating sequences regulate targets of the unfolded protein response. *PLoS Biol* 2, E246.

Peckol, E. L., Troemel, E. R., and Bargmann, C. I. (2001). Sensory experience and sensory activity regulate chemosensory receptor gene expression in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* *98*, 11032-11038.

Portman, D. S., and Emmons, S. W. (2000). The basic helix-loop-helix transcription factors LIN-32 and HLH-2 function together in multiple steps of a *C. elegans* neuronal sublineage. *Development* *127*, 5415-5426.

Robertson, H. M. (1998). Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res* *8*, 449-463.

Robertson, H. M. (2000). The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. *Genome Res* *10*, 192-203.

Sagasti, A., Hobert, O., Troemel, E. R., Ruvkun, G., and Bargmann, C. I. (1999). Alternative olfactory neuron fates are specified by the LIM homeobox gene *lim-4*. *Genes Dev* *13*, 1794-1806.

Sarafi-Reinach, T. R., Melkman, T., Hobert, O., and Sengupta, P. (2001). The *lin-11* LIM homeobox gene specifies olfactory and chemosensory neuron fates in *C. elegans*. *Development* *128*, 3269-3281.

Sarafi-Reinach, T. R., and Sengupta, P. (2000). The forkhead domain gene *unc-130* generates chemosensory neuron diversity in *C. elegans*. *Genes Dev* 14, 2472-2485.

Satterlee, J. S., Sasakura, H., Kuhara, A., Berkeley, M., Mori, I., and Sengupta, P. (2001). Specification of thermosensory neuron fate in *C. elegans* requires *ttx-1*, a homolog of *otd/Otx*. *Neuron* 31, 943-956.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J Mol Biol* 188, 415-431.

Sengupta, P., Chou, J. H., and Bargmann, C. I. (1996). *odr-10* encodes a seven transmembrane domain olfactory receptor required for responses to the odorant diacetyl. *Cell* 84, 899-909.

Sengupta, P., Colbert, H. A., and Bargmann, C. I. (1994). The *C. elegans* gene *odr-7* encodes an olfactory-specific member of the nuclear receptor superfamily. *Cell* 79, 971-980.

Shimizu, T., Toumoto, A., Ihara, K., Shimizu, M., Kyogoku, Y., Ogawa, N., Oshima, Y., and Hakoshima, T. (1997). Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *Embo J* 16, 4689-4697.

Troemel, E. R., Chou, J. H., Dwyer, N. D., Colbert, H. A., and Bargmann, C. I. (1995). Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans*. *Cell* 83, 207-218.

Troemel, E. R., Kimmel, B. E., and Bargmann, C. I. (1997). Reprogramming chemotaxis responses: sensory neurons define olfactory preferences in *C. elegans*. *Cell* 91, 161-169.

Wes, P. D., and Bargmann, C. I. (2001). *C. elegans* odour discrimination requires asymmetric diversity in olfactory neurons. *Nature* 410, 698-701.

Zheng, P., Pennacchio, L. A., Le Goff, W., Rubin, E. M., and Smith, J. D. (2004). Identification of a novel enhancer of brain expression near the apoE gene cluster by comparative genomics. *Biochim Biophys Acta* 1676, 41-50.

Table 1. Distribution of word lengths and their occurrences in a probabilistic segmentation of the upstream sequences of 920 *C. elegans* chemoreceptor genes.

Length	Words in dictionary	Appearances in ML partition	Sequence explained
1	4	356053	57.54%
2	3	14674	4.74%
3	17	20438	9.91%
4	75	15694	10.14%
5	90	12019	9.71%
6	86	4347	4.21%
7	74	1425	1.61%
8	128	1145	1.48%
9	77	304	0.44%
10	31	92	0.15%
11	8	32	0.06%

Table 2. Sequence motifs with statistically significant positional preference in the promoters of *C. elegans* chemosensory receptor genes.

Word	-1000 to -900	-900 to -800	-800 to -700	-700 to -600	-600 to -500	-500 to -400	-400 to -300	-300 to -200	-200 to -100	-100 to ATG	χ^2	P	Comment
tcacctg	9	4	7	5	6	4	6	13	20	43	111.8	2.3×10^{-19}	E-box
caggtgaaa	1	2	2	1	4	1	1	2	7	24	101.0	3.4×10^{-17}	E-box
cacctgc	6	3	4	4	3	2	5	7	16	33	98.8	9.5×10^{-17}	E-box
acaggtgaa	1	1	1	0	0	1	0	3	5	13	57.8	9.4×10^{-9}	E-box
agcagctgaaa	0	0	1	0	0	1	0	0	0	8	56.0	2.1×10^{-8}	E-box
gcaggtgaa	0	4	0	1	2	0	1	0	5	12	51.4	1.5×10^{-7}	E-box
gtctag	12	10	20	13	13	10	11	15	26	37	41.0	1.1×10^{-5}	SMAD
gcaggtg	6	7	3	2	9	9	4	4	7	22	40.0	1.7×10^{-5}	E-box
tccacctgtt	0	0	0	1	0	1	0	0	0	6	39.5	2.1×10^{-5}	E-box
cacctgtc	2	1	5	0	1	1	1	4	7	11	33.4	2.3×10^{-4}	E-box
ctagac	12	12	21	16	14	13	22	15	37	27	30.9	6.1×10^{-4}	SMAD
ctataatt	1	6	4	2	5	12	8	8	10	17	28.8	1.3×10^{-3}	CdxA

Table 3. Functional preference of candidate regulatory motifs for functionally defined families of *C. elegans* genes. Shown are the Gene Ontology categories in which motifs are significantly overrepresented in the proximal promoter sequences.

Motif	Gene Ontology (GO) category	Genes in GO set	Genes in GO set with motif	Over-representation factor	P
cacctg	G-protein-coupled receptor Nucleosome	807	92	4.08	$P < 10^{-28}$
		79	12	5.44	$P < 10^{-5}$
cagctg	G-protein-coupled receptor Nucleosome	807	32	2.57	$P < 10^{-5}$
		79	11	9.04	$P < 10^{-7}$
ctagac	G-protein-coupled receptor	807	74	2.87	$P < 10^{-14}$
ctataatt	G-protein-coupled receptor	807	9	2.28	$P < 10^{-3}$

Table 4. Nucleotide frequencies at positions surrounding appearance of the asymmetric E-box core CACCTG in the proximal promoters of *C. elegans* chemoreceptor genes.

a	c	g	t	Bits	Chi-square	Cons
64	26	17	36	0.05	16.43	
75	12	14	42	0.13	41.13	
54	10	20	59	0.09	23.60	
48	7	3	85	0.30	81.15	T/A
37	7	8	91	0.28	85.89	T/A
12	23	15	93	0.28	88.33	T/C
	143			1.48	572.00	C
143				1.09	333.67	A
	143			1.48	572.00	C
	143			1.48	572.00	C
			143	1.09	333.67	T
		143		1.48	572.00	G
11	59	13	60	0.25	71.36	T/C
20	38	15	70	0.13	38.90	T/C
44	34	23	42	0.01	2.16	
32	38	23	50	0.03	8.13	
30	24	31	58	0.03	10.14	
37	27	32	47	0.01	1.70	

Table 5. Expression patterns of chemosensory receptor genes with proximal ebox motifs, assayed by promoter::gfp fusion. In each case, 3-4 kb of upstream sequence was fused to the reading frame of gfp. The score shown measures the strength of match to the consensus sequence identified in Table 5.

<u>Gene</u>	<u>Position</u>	<u>Strand</u>	<u>Sequence</u>	<u>Score</u>	<u>Expression</u>
<i>srh-220</i> F47C12.5	-96	+	catttcacctgctgcgt	8.72	ADL *
<i>srh-186</i> F36G9.2	-70	-	tatttcagctgctcagt	8.72	ADL
<i>srh-132</i> T27C5.5	-58	-	aaatttcacctgtgcaa	8.69	ADL, midbody
<i>sri-51</i> ZC239.8	-104	+	tgtttcacctgttttg	8.40	ADL, AIN, PHA, PHB
<i>sro-1</i> DI1022.6	-66	-	tgttccacctgtttcta	8.36	ADL, SIA *
<i>srh-169</i> C49G7.2	-68	-	tgtttcagctgctgttc	8.29	none detected *
<i>srh-204</i> E03D2.3	-74	-	aaaaatcacctgttccaa	8.06	AWC, AWA, ASI *
<i>str-118</i> F57A8.3	-66	-	aattatcagggtgcttcag	7.34	AVJ or AIN *
<i>srh-213</i> T22F3.5	-57	-	actttacacctgttatgt	7.72	Interneurons, muscle *

* From Troemel et al., 1995.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

Table 6. Alignment of the basic DNA-binding domains of predicted bHLH transcription factors in the *C. elegans* genome. “+” indicates the presence of a large, non-polar residue at position 13, a requirement for binding the DNA sequence cagctc rather than cacgtg. “*” indicates the presence of a small, polar residue at position 3, a requirement for binding E-boxes with a thymine nucleotide on the 5’ side of the E-box core. Only four of the *C. elegans* bHLH proteins – *hlh-2*, *lin-32*, *hlh-12*, and *hlh-15* – appear to be likely candidates for binding the “tttcacstg” motif shared by the ADL-expressed genes (Table 6).

	1234567890123	
<i>hlh-1</i>	RRKAATMRERRRL	+
<i>hlh-2</i>	RRSQNNARERVV	+*
<i>hlh-3</i>	TKQKRNERERKRV	+
<i>hlh-4</i>	VVAKRNARERTRV	+
<i>hlh-5</i>	RRVKANGRERARM	+
<i>hlh-6</i>	SVWKRNERERCRV	+
<i>hlh-7/lin-32</i>	RRSAANERERRRM	+*
<i>ngn-1</i>	RRDKANARERRRM	+*
<i>cnd-1</i>	RRVKANGRERARM	+
<i>hlh-8</i>	QRACANRRERQRT	
<i>hlh-9</i>	KIKNKPLMEKKRR	
<i>hlh-10</i>	RRYEANARERNRV	+
<i>hlh-11</i>	RRQIANCNERRRM	+
<i>hlh-12</i>	RRSRANERERQRV	+*
<i>hlh-13</i>	ERQTASIRERKRM	+
<i>hlh-14</i>	KEAMAKKNQVARN	
<i>hlh-15</i>	YRNLHATRERIRV	+
<i>hlh-16</i>	IKQLANANHKLQM	+
<i>hlh-17</i>	VRLSINLRERCRM	+
<i>hlh-18/hnd-1</i>	SKKSrKEKSREKE	
<i>hlh-19</i>	PSKAETLKSAAQY	
<i>hlh-20</i>	RRTAHNLIEKKYR	
<i>hlh-22/hnd-1</i>	SKKSrKEKSREKE	
<i>hlh-23/mxl-3</i>	RRAHHNELEERRR	
<i>hlh-24/ref-1</i>	RKEVKKNREQDRR	
<i>hlh-25</i>	RRKVKTEREKIRR	
<i>hlh-26</i>	IKKIKSDREQVRR	
<i>hlh-27</i>	RRKVKTEREKIRR	
<i>hlh-28</i>	KQKVKTKREQIRR	
<i>hlh-29</i>	KQKVKTKREQIRR	

Figure 1. Positional preference of candidate regulatory motifs and related sequences in *C. elegans* chemoreceptor gene promoters.

Figure. 1A Positional preference in *C. elegans* chemoreceptor gene promoters of the E-box motifs cacctg, caggtg, and cagctg, all of which were identified by probabilistic segmentation analysis; and, for comparison, of the E-box sequence cacgtg, which was not identified. bHLH transcription factors readily distinguish between cacgtg and the other sequences.

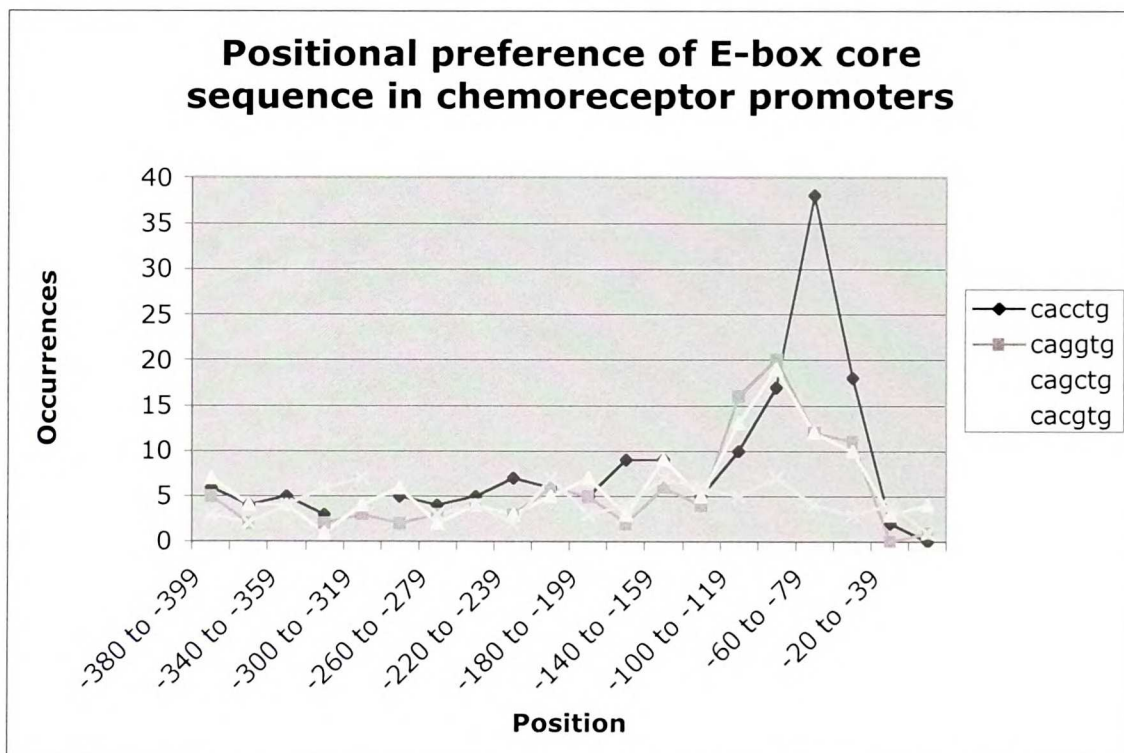


Figure 1B Positional preference in *C. elegans* chemoreceptor promoters of the SMAD-binding motifs ctagac and gtctag, which were identified by probabilistic segmentation analysis.

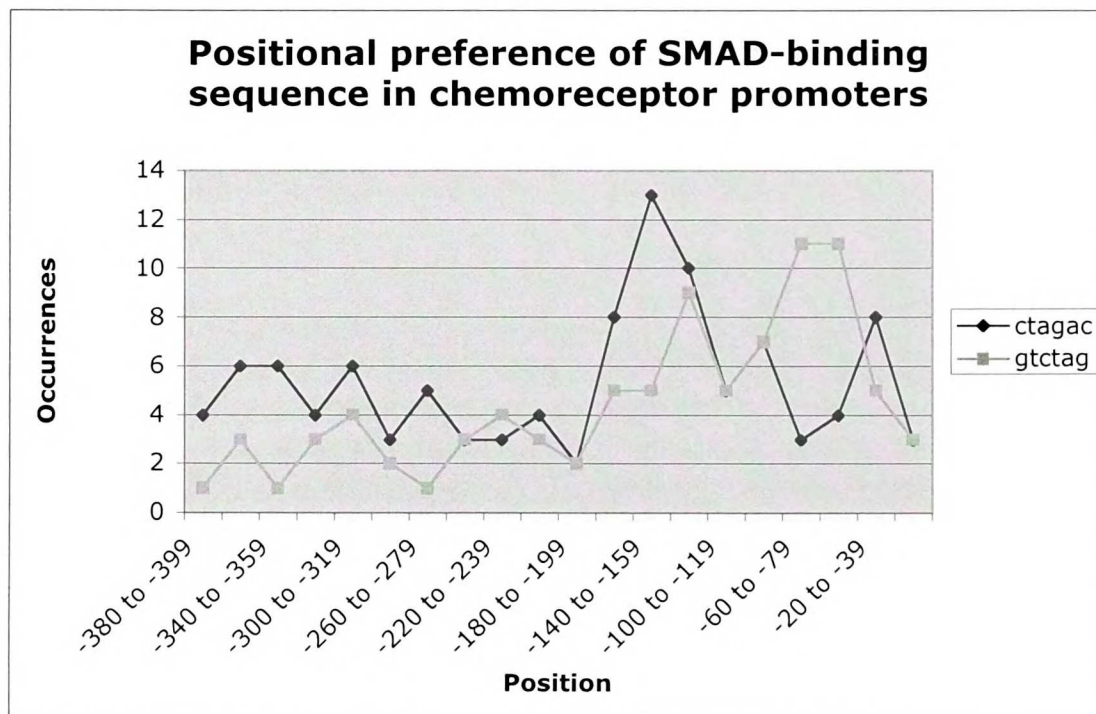


Figure 1C. Positional preference and strand bias of the CdxA-binding motifs ctataatt, which was identified by probabilistic segmentation analysis, versus its inverse complement sequence aattatag, which was not.

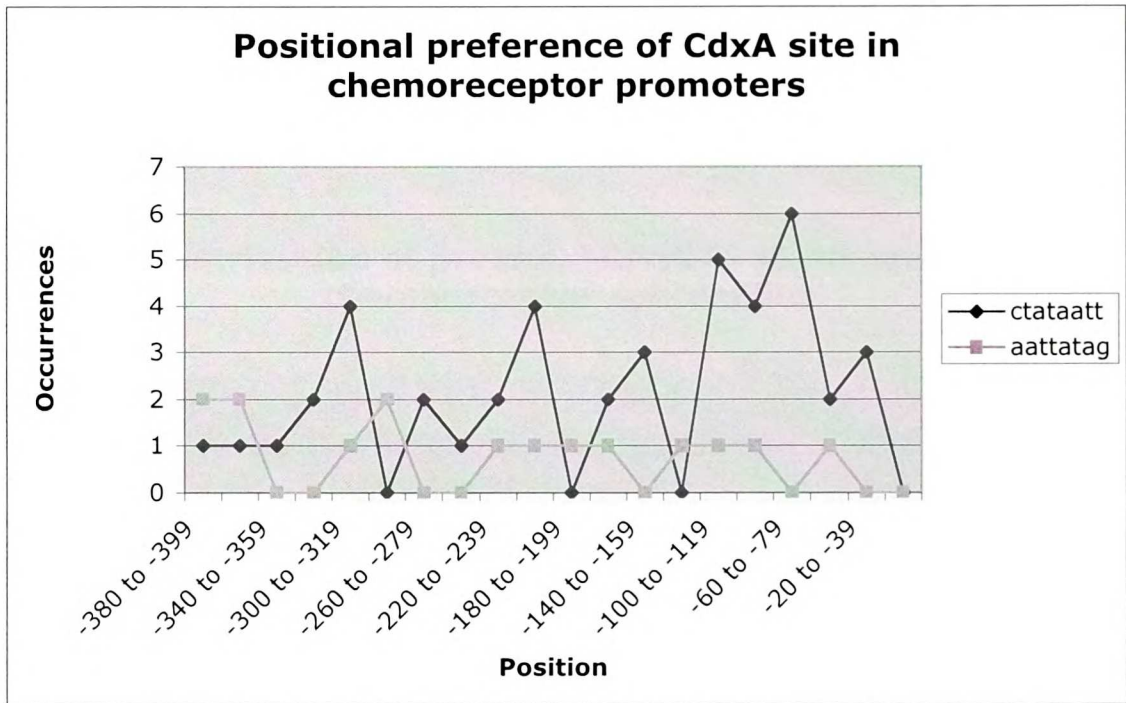
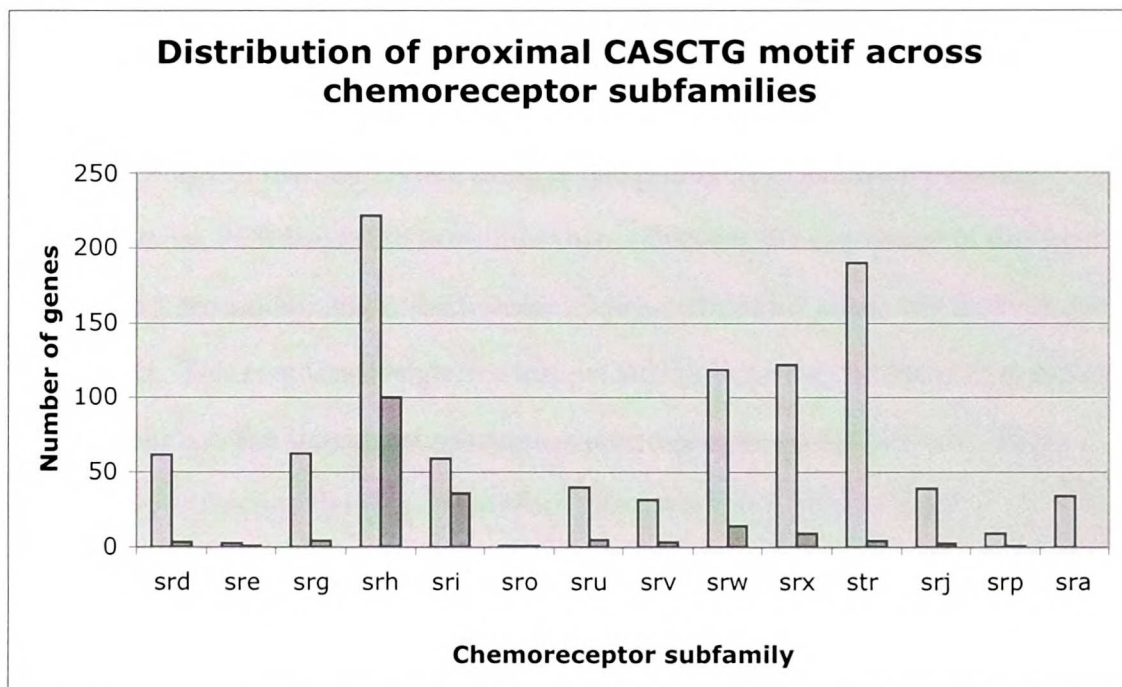


Figure 2. Distribution of proximal CASCTG E-box motifs in the proximal promoters of *C. elegans* chemoreceptor genes of different sequence families. Light bars: Total number of genes in family. Dark bars: Number of genes with CASCTG motifs in proximal promoter. Background frequency in the genome (across the proximal promoters of all genes) is 2%.



RY
UNIVERSIT
UNIVERSIT
81
un
UNIVERSIT
UNIVERSIT
81
un
-ALFOR
-ALFOR
ncis
RY
UNIVERSIT
UNIVERSIT
81
un

18
un
UNIVERSIT
UNIVERSIT
81
un

APPENDIX

A USEFUL MODULAR SYSTEM FOR MAKING C. *ELEGANS* EXPRESSION CONSTRUCTS

C. elegans researchers routinely make expression constructs in which an open reading frame of interest is expressed under the control of a promoter sequence from a *C. elegans* gene. These constructs are typically designed by identifying unique, compatible restriction sites in the vector, promoter, and open reading frame of interest. When unique, compatible restriction sites do not already exist, PCR is used to introduce them. Because the sequences of different genes and promoters differ, the usable unique, compatible sites vary from project to project. This requires design of a unique subcloning strategy for each project and minimizes the sharing of subcloning intermediates across projects. These project-specific considerations introduce extra work and unnecessary heterogeneity into constructs used by *C. elegans* researchers.

The pSM cloning vector is designed to be a flexible, modular *C. elegans* expression vector that will allow almost any *C. elegans* gene promoter to be efficiently cloned into the same restriction sites and rapidly exchanged across expression constructs. Once a promoter is amplified and cloned into any pSM construct, it can readily be inserted into any other pSM construct using the same restriction sites, without the need for a construct-specific subcloning strategy or the introduction of alternative subcloning sites by additional PCR.

pSM is derived from the expression vector pPD49.27. It contains two multiple cloning sites (one for promoters, one for ORFs) separated by a synthetic intron (to enhance gene expression), and followed by the *unc-43* 3'UTR (also to

LIBRARY
UNIVERSITY
OF
ALBANY
STATE UNIVERSITY OF NEW YORK
LIBRARY
UNIVERSITY
OF
ALBANY
STATE UNIVERSITY OF NEW YORK
LIBRARY
UNIVERSITY
OF
ALBANY
STATE UNIVERSITY OF NEW YORK

LIBRARY
UNIVERSITY
OF
ALBANY
STATE UNIVERSITY OF NEW YORK
LIBRARY
UNIVERSITY
OF
ALBANY
STATE UNIVERSITY OF NEW YORK

increase expression) downstream of the second multiple cloning site. We introduced into the first multiple cloning site a series of restriction sites – in particular, NotI, FseI, and AscI -- that would be of maximal utility to *C. elegans* researchers, given the distribution of restriction sites in the *C. elegans* genome.

FseI and AscI were chosen because:

1. FseI and AscI cut very rarely in the *C. elegans* genome. Due to the genome's AT-richness and the GC-richness of the FseI (GGCCGGCC) and AscI (GGCGCGCC) restriction sites, the *C. elegans* genome has only 231 FseI sites and 202 AscI sites. Fewer than 1% of *C. elegans* genes have either site in their upstream (3 kb) sequence.
2. FseI and AscI leave ends which are incompatible for accidental religation – a 5' AscI overhang and a 3' FseI overhang -- minimizing background religation in the absence of insert.
3. FseI and AscI leave “sticky” 4-base GC-rich overhangs, maximizing the efficiency of ligation to compatible sequences on insert DNA.
4. FseI and AscI are inexpensive and cut efficiently in the same reaction buffer.

In addition, a NotI site was added upstream of the FseI site, to allow researchers to subsequently insert additional, farther-upstream promoter segments via directional cloning, after an initial expression construct has been created.

The most important feature of the FseI and AscI restriction sites is their scarcity in the *C. elegans* genome, which allows them to be used for directional cloning of almost all *C. elegans* promoters. This allows:

ERSIT
ERSIT
BI
an
ERSIT
ERSIT
BI
an
AUFOR
AUFOR
ncis
RI
IVERSIT
IVERSI
BI
an

LIBRARY
1951

1. Facile sharing of subcloning intermediates -- such as promoters into which flanking FseI and AscI sites have been introduced -- among projects, via simple cut-and-paste subcloning.
2. Efficient parallel processing of subcloning reactions in which a single ORF is placed under the control of many different promoters.
3. Reaction multiplexing. For example, the cloning of a large number of promoters, into which flanking FseI and AscI sites have been introduced, into a common vector via a single ligation / cloning reaction. The recombinant clones can subsequently be distinguished by re-sequencing or a restriction digest.

Maps of pSM and pSM/gfp have been deposited in my notebook and with the lab.



100
20

UNIVERSITY

UNIVERSITY

UNIVERSITY

18
100

7374880



3 1378 00737 4880

